V. Dulin

ELECTRON DEVICES

Mir Publishers Moscow

V. Dulin

ELECTRON DEVICES

Viktor N. Dulin, Cand. Tech. Sc., Associate Professor, Chair of theoretical radio engineering at the Moscow Aviation Institute, author of several books and a number of papers, member of various educational and scientific councils.

V. Dulin

ELECTRON DEVICES

ИЗДАТЕЛЬСТВО
МИР

В. Н. Дулин

ЭЛЕКТРОННЫЕ ПРИБОРЫ

Издательство
«ЭНЕРГИЯ»
Москва

V. Dulin

ELECTRON DEVICES

Translated from Russian
by
A. KUZNETSOV

MIR
PUBLISHERS
MOSCOW

First published 1980

Revised from the 1977 Russian edition

The Greek Alphabet

Α α	Alpha	Ι ι	Iota	Ρ ρ	Rho
Β β	Beta	Κ κ	Kappa	Σ σ	Sigma
Γ γ	Gamma	Λ λ	Lambda	Τ τ	Tau
Δ δ	Delta	Μ μ	Mu	Υ υ	Upsilon
Ε ε	Epsilon	Ν ν	Nu	Φ φ	Phi
Ζ ζ	Zeta	Ξ ξ	Xi	Χ χ	Chi
Η η	Eta	Ο ο	Omicron	Ψ ψ	Psi
Θ θ ϑ	Theta	Π π	Pi	Ω ω	Omega

The Russian Alphabet and Transliteration

А а	a	К к	k	Х х	kh
Б б	b	Л л	l	Ц ц	ts
В в	v	М м	m	Ч ч	ch
Г г	g	Н н	n	Ш ш	sh
Д д	d	О о	o	Щ щ	shch
Е е	e	П п	p	Ъ ъ	″
Ё ё	yo	Р р	r	Ы ы	y
Ж ж	zh	С с	s	Ь ь	′
З з	z	Т т	t	Э э	e
И и	i	У у	u	Ю ю	yu
Й й	y	Ф ф	f	Я я	ya

На английском языке

© Издательство «Энергия», 1977
© English translation, Mir Publishers, 1980

CONTENTS

Preface 9

PART ONE. ELECTRON TUBES

Chapter One. **A General Outline and Physical Principles of Electron Tubes** 11

1-1. Classification and Design of Electron Tubes 11
1-2. Operating Conditions, Ratings and Characteristics of Electron Tubes 14
1-3. Electron Emission 16
1-4. Thermionic Cathodes 21
1-5. Electric Current in a Vacuum 26

Chapter Two. **Vacuum Diodes** 29

2-1. Physical Processes in the Vacuum Diode 29
2-2. Plate Current as a Function of Plate Voltage 32
2-3. Static Characteristics of the Vacuum Diode 35
2-4. The Load Line 37
2-5. Parameters and Ratings of the Vacuum Diode 49
2-6. Construction and Types of Vacuum Diodes 43

Chapter Three. **Vacuum Triodes** 45

3-1. Physical Processes in the Vacuum Triode 45
3-2. Static Characteristis of the Vacuum Triode 53
3-3. Parameters and Ratings of the Vacuum Triode 59
3-4. Operation of the Vacuum Triode Under Load 67
3-5. Construction and Types of Vacuum Triodes 73

Chapter Four. **Multielectrode and Special-Purpose Tubes** 76

4-1. Tetrodes and Pentodes. Physical Processes in Multielectrode Tubes 76
4-2. Static Characteristics of Multielectrode Tubes 85
4-3. Parameters of Multielectrode Tubes 89

4-4. Construction and Types of Tetrodes and Pentodes 95
4-5. Vacuum Tubes for Wideband Amplifiers 97
5-6. Radio-Frequency Tubes 100
4-7. Special-Purpose Tubes 107

Chapter Five. **Operational Hints for Vacuum Tubes** 115

5-1. Tube Noise 115
5-2. Equivalent Circuits of Vacuum Tubes 121
5-3. Approximation of Tube Characteristics 124
5-4. Reliability of Vacuum Tubes. Operating Conditions and Ratings 129

Chapter Six. **Photoelectric Devices** 131

6-1. General 131
6-2. Photoelectric Emission 133
6-3. Phototubes 134
6-4. Photomultiplier Tubes 138

Chapter Seven. **Electron-Beam Devices** 141

7-1. Construction and Principle of Operation of Cathode-Ray Tubes 141
7-2. Intensity Modulation of the Beam 143
7-3. Focusing Structures. The Electron Gun 146
7-4. Deflection Structures 153
7-5. CRT Screens 158
7-6. Construction and Types of Cathode-Ray Tubes 162

Chapter Eight. **Gas-Filled Tubes** 170

8-1. General 170
8-2. Electric Discharge in Gas 170
8-3. Arc-Discharge Tubes 176
8-4. Glow- and High-Frequency-Discharge Tubes 179

PART TWO. SEMICONDUCTOR DEVICES

Chapter Nine. **A General Outline and Physical Principles of Semiconductor Devices** 185

9-1. Classification and Design of Semiconductor Devices 185
9-2. Energy Diagrams of Semiconductors 189
9-3. Generation and Recombination of Free Charge Carriers in Semiconductors 201
9-4. Particle Statistics as Applied to Semiconductors 204
9-5. Concentration of Free Charge Carriers in Semiconductors . . . 210
9-6. Motion of Charge Carriers. Electrical Conduction in Semiconductors 216

Chapter Ten. **Junction Effects** 226

10-1. Semiconductor Junctions 226
10-2. The Symmetrical P-N Junction 226
10-3. Forward and Reverse Biasing 231
10-4. Types of Junctions 238

10-5. Breakdown of the *P-N* Junction 243
10-6. Capacitances of the *P-N* Junction 245

Chapter Eleven. **Crystal Diodes** 249

11-1. Design and Classification of Crystal Diodes 249
11-2. Physical Processes in the Crystal Diode 250
11-3. The Volt-Ampere Characteristic of the Real Diode 258
11-4. The Main Parameters of Crystal Diodes 261
11-5. Crystal Diode Transients 263
11-6. Rectifier Diodes 267
11-7. Crystal Diodes for Pulse Work 268
11-8. Mixer and Detector Diodes 271
11-9. Zener Diodes 273
11-10. Varicaps 275

Chapter Twelve. **Bipolar Transistors** 278

12-1. Design and Principle of Operation 278
12-2. Physical Processes in the Transistor and Its Parameters . . . 285
12-3. The Transistor as a Two-Port 296
12-4. Static (D.C.) Characteristics of the Transistor 298
12-5. Static (D. C.) Transistor Parameters 310
12-6. Small-Signal Operation of a Transistor 312
12-7. R.F. and Pulse Operation of the Transistor 320
12-8. Effect of the Operation Mode and Temperature on Transistor Characteristics and Parameters 329
12-9. Construction and Types of Bipolar Transistors 333

Chapter Thirteen. **Field-Effect Transistors** 339

13-1. General 339
13-2. Junction Field-Effect Transistors 341
13-3. Insulated-Gate Field-Effect Transistors 345
13-4. FET Parameters 352

Chapter Fourteen **Photoelectric and Light-Emitting Semiconductor Devices** 355

14-1. General 355
14-2. Absorption of Light by Semiconductors 355
14-3. The Photoconductive Effect 359
14-4. Photoresistors 361
14-5. The Photovoltaic Effect 365
14-6. Photoelectric Cells 369
14-7. Photodiodes 374
14-8. Phototransistors 381
14-9. Light-Emitting Semiconductor Devices 384

Chapter Fifteen **Miscellaneous Semiconductor Devices** 389

15-1. Negative Resistance Devices 389
15-2. The Unijunction Transistor 390
15-3. Thyristors 391
15-4. The Tunnel Diode 396

Chapter Sixteen. **Applications of Semiconductor Devices** 403

16-1. Noise in Semiconductor Devices 403
16-2. Linear Approximation of Diode Characteristics and Diode Equivalent Circuits 407
16-3. Equivalent Circuits for Bipolar Transistors 409
16-4. Linear Approximation of Characteristics for a Bipolar Transistor 412
16-5. Equivalent Circuits of Field Effect Transistors 413
16-6. Reliability of Semiconductor Devices and Operating Variables 415

References 418

Index 420

PREFACE

This book has been written to cover the course in electron devices as taught in the Soviet Union to students majoring in electronic engineering.

The course includes basic theory, design, principles of operation, characteristics, parameters and application of most present-day electron tubes and semiconductor devices.

Part One of the book deals with electron tubes. It is devoted to the physical principles of electron tube operation (electron emission, cathode parameters, current in a vacuum, etc.), main physical processes in vacuum, cathode-ray and gas-filled tubes, their main characteristics and parameters.

Part Two discusses semiconductor devices. The beginning of this part presents a survey of the most important processes in semiconductor devices: the generation and recombination of charge carriers, the laws describing their motion in solids, and also the processes taking place in transition regions.

The rest of Part Two deals with the principles of operation, characteristics and parameters of various diodes, bipolar and field-effect transistors, thyristors, etc.

Special emphasis is placed on photoelectric and light-emitting devices, such as photoresistors, photodiodes, phototransistors, light-emitting diodes, and so on.

In addition, attention is centred on application of various electron devices, noise, equivalent circuits, approximation of their characteristics, reliability, etc.

All comments from the reviewers were accepted with gratitude and used in work on the manuscript.

The author wishes to thank Cands. Tech. Sc. N. Alekseev, Yu. Naumov, G. Shishkin and N. Fedorov for their valuable suggestions, and T. Kozlova and A. Belskaya for their assistance in the preparation of the manuscript.

Part One

ELECTRON TUBES

Chapter One

A GENERAL OUTLINE AND PHYSICAL PRINCIPLES OF ELECTRON TUBES

1-1. CLASSIFICATION AND DESIGN OF ELECTRON TUBES

Definition. *An electron tube* is an electron device in which conduction by electrons takes place through a vacuum or gaseous medium within a gastight envelope.

Classification. According to the medium through which electrically charged particles move, electron tubes may be classed into *vacuum* and *gas-filled* types.

A *vacuum tube* is a device in which conduction is only by free electrons moving through a high, or hard, vacuum.

A *gas-filled tube* (or, simply, gas tube) is one in which conduction is owing to electric discharge in gas or vapour.

Vacuum electron devices are many and diverse. They include vacuum tubes proper, electron-beam devices, and vacuum photoelectric devices.

Vacuum tubes are intended for conversion of electrical quantities. They make up the most representative group among electron tubes. They may be classed by the number of electrodes as diodes (with two electrodes), triodes (with three electrodes), or multielectrode tubes, such as tetrodes, pentodes, hexodes, heptodes and octodes (those with four, five, six, seven and eight electrodes, respectively). According to the functions they perform, vacuum tubes may be classified as generating, amplifying, rectifying, frequency-conversion, and so on. Vacuum tubes are specifically designed for operation at audio frequencies, or at radio frequencies, or in the microwave region. They may also be classed by the envelope material, base type, cathode type or some other characteristic.

An *electron-beam device* is an electron tube which depends for its operation on the formation and control of one or more electron beams. By the type of control they may be divided into electrostatically- and magnetically-controlled devices, and, in accordance with the number of electron beams, they may be classed as single-beam, double-beam and multibeam tubes. Electron-beam devices include

cathode-ray tubes used in oscilloscopes, radars, television receivers, etc.

Still another variety of electron tubes are *photoelectric devices* whose electrical properties vary under the action of light. They include *phototubes* — electron tubes containing photocathodes, and *multiplier phototubes* — phototubes in which photoemission current is multiplied owing to secondary emission. There are also *TV camera*

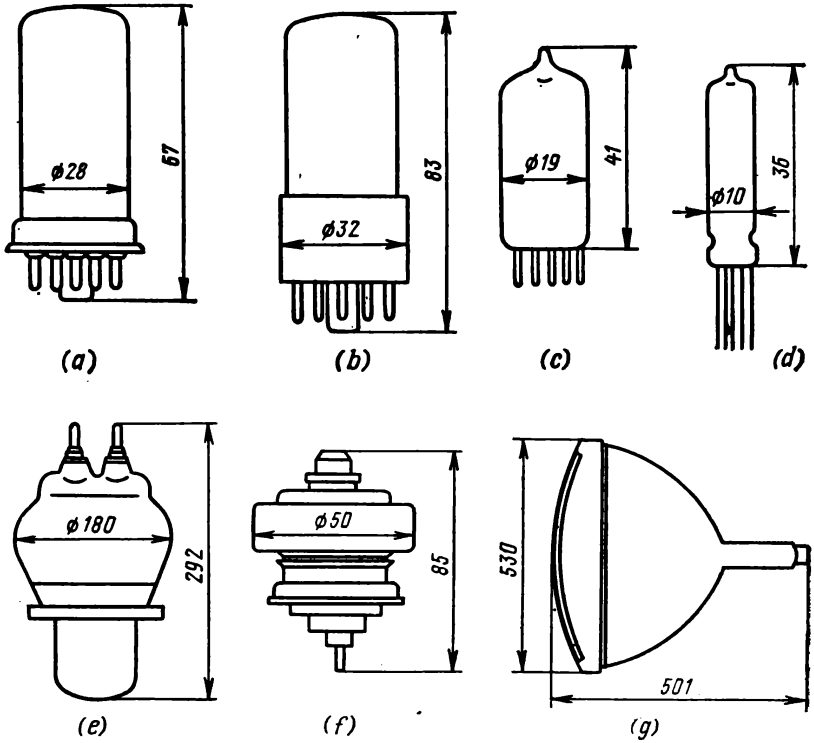


Fig. 1-1. Electron tube envelopes
(*a*) metal envelope; (*b*) conventional glass envelope; (*c*) miniature glass envelope; (*d*) micro-miniature glass envelope; (*e*) metal-glass envelope; (*f*) metal-ceramic envelope; (g) TV picture tube (kinescope) envelope

tubes in which an electron current or charge-density image is formed from an optical image and scanned to provide an electric signal.

Gas-filled electron devices are classed by the type of discharge (arc-discharge, glow-discharge, etc.), by their purpose (VR tubes, gas rectifiers, etc.), and also by their design.

Design. As already noted, electric processes in an electron tube take place within an *envelope*. The envelopes of vacuum tubes are evacuated to a pressure of 10^{-5} to 10^{-4} Pa. Those of gas tubes are filled with inert gases (such as argon, neon, krypton or their mixtures), hydrogen or mercury vapour at a pressure of 10^{-1} to 10^{2} Pa.

The electric processes taking place within the envelope, may be visualized as follows: free electric charge carriers are formed, such as electrons emitted from the surface of a solid body or electrons and ions produced by electric discharge in a gas; these charge carriers form a current controlled in density and direction. All these processes are carried out by *electrodes* (cathodes, grids, anodes, plates, etc.) connected to voltage sources and external circuits by pins which are often combined by a part known as a *base*.

The envelopes of electron tubes differ in size, shape and material (Fig. 1-1). The most widely used are cylindrical envelopes, though use is sometimes made of envelopes having a more complex shape. The overall dimensions of envelopes range from several millimetres (microminiature tubes) to tens of centimetres (high-power transmitting tubes and electron-beam devices). Electron tubes may have envelopes made of special glass which can easily be degassed and joined to metals to form an airtight seal. They may have metal or metal-glass envelopes. There are also r.f. ceramic envelopes manufactured in combination with metal parts.

The simplest electron tube (diode) contains only two electrodes, namely the cathode and the anode (or plate). More complex devices have also grids and other electrodes.

The *cathode* is an electrode intended for emission of free electrons. Most electron tubes (vacuum, cathode-ray and some gas-filled tubes) use *thermionic*, or *hot*, *cathodes* which supply electrons by thermionic emission, that is, due to the rise in their temperature. Some tubes use *cold cathodes* and *secondary-emission cathodes* (such as *dynodes*). The former emit electrons under the action of an electric field, the latter supply free electrons when struck by other electrons. There are also *photocathodes* which release electrons when exposed to light or other suitable radiation.

The *anode* (or plate) is the electrode to which the principal stream of electrons flows.

The *grids* are electrodes designed to establish an electric field within the envelope of an electron tube to control the electron flow.

Cathodes, grids and anodes are made in many types according to their purpose and operating conditions.

Electrode leads in electron tubes also differ in construction. The most commonly used are base-type electron tubes (Fig. 1-2*a*). Here, the electrode leads made of soft wire are sealed in a glass disc, known as a stem, somewhat thicker than the walls of the envelope, which is then sealed to the envelope. The wire leads are welded to the pins press-fitted into the bottom of a plastic base which is cemented to the glass envelope. Figure 1-2*a* shows an octal base which contains eight pins arranged in a symmetrical pattern round the circumference. Located in the centre of the base is an alignment pin with a rib, which provides for the proper insertion of the tube in its socket.

Figure 1-2*b* and *c* shows a stiff- and a soft-lead tube which have no base. In them, stiff or soft wire leads are sealed in the stem or the

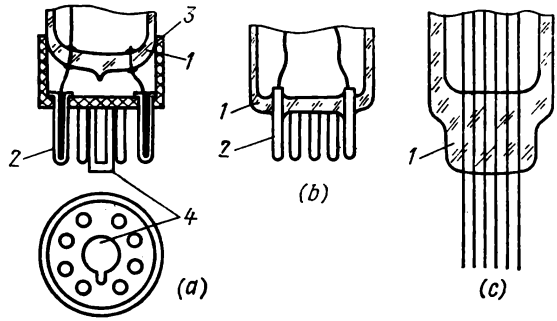


Fig. 1-2. Electrode leads and pin connections
(*a*) pins in base; (*b*) stiff-lead construction; (c) soft-lead construction; *1*—stem; *2*—pins; *3*—plastic base; *4*—alignment pin

flattened end of the envelope. In some tubes, the anode (plate) or grid leads are brought out through the glass envelope on the top or side rather than through the base.

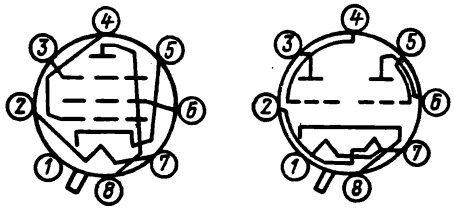


Fig. 1-3. Base (pin-connection) diagrams of electron tubes

The diagram showing the connection of the tube electrodes to the pins is called a *base* (or *pin-connection*) *diagram*. Base diagrams for a particular class of tubes are always the same and can be found in tube manuals. Examples are shown in Fig. 1-3.

Diagram symbols and type designations. In the Soviet Union, electron tubes are symbolized in circuit diagrams in accordance with a USSR State Standard, GOST 7624-62. The system for marking electron tubes is set up by GOST 13393-67.

1-2. OPERATING CONDITIONS, RATINGS AND CHARACTERISTICS OF ELECTRON TUBES

An analysis of electron-tube operation requires knowledge of their operating conditions, ratings and characteristics.

Electrode Voltages. In electron tubes, the flow of electrons emitted by the cathode is controlled by electric fields set up by the electrodes whose potentials are maintained by external sources. Unless otherwise specified, the cathode potential is taken to be zero, $V_k = 0$, and the potentials at all other electrodes are determined relative to that level. The potential difference between an electrode and the

cathode bears the name of that electrode. Since $V_k = 0$, the voltage at a particular electrode is numerically equal to its potential.

Electron Tube Currents. The current which flows in the external circuit of a particular electrode bears the name of that electrode; for example, *anode* (or *plate*) *current*, *grid current*, etc.

Operating Conditions. These are the conditions to which a tube is subjected during operation or quantities that characterize the work of a tube. For example, *electrical operating conditions* characterize electrode voltages and currents. *Mechanical operating conditions* refer to mechanical factors which act upon an operating tube (such as shocks and jarring). The working temperature range, relative humidity and other similar quantities are known as *climatic operating conditions*. Each quantity making up the operating conditions is called a *variable*.

The electrodes of electron tubes may be connected to sources of unvarying and varying voltages. If the voltages at all electrodes of a tube are constant, the tube is said to operate in a *static mode*. In this case, all quantities defining the operating conditions remain unchanged with time.

If at least one of them varies with time, the tube is said to operate in a *dynamic mode*. In the specific case, when the quantities characterizing the work of a tube vary so slowly that its operating conditions at any time differ insignificantly from the static operation (that is, its operation obeys static laws), the tube is said to operate in a *quasi-static mode*. In the general case, the variables of the dynamic operation are interrelated differently from the static mode.

The electrode leads of electron tubes used in various electronic circuits may contain resistors, resonant circuits, coils, and so on. These are called *load elements*, or, simply, *load*.

A tube containing a load in at least one electrode lead is said to operate *under load*.

Electron Tube Ratings. Tube manufacturers, standards or specifications establish safe and limiting capabilities or conditions under which the tubes may be operated. These include nominal (or rated) filament voltage, maximum electrode currents, maximum anode (plate) dissipation, interelectrode capacitances, and so on, and are called *tube ratings*.

Electron Tube Characteristics. The relationship between a pair of tube variables, with all other variables held constant, is called a *tube characteristic*.

Two types of electron tube characteristics are distinguished, *static* and *dynamic*. The static characteristics represent the relationship between the voltages at one or two tube electrodes and an electrode current under no-load (static) conditions, with all other voltages maintained constant. The dynamic characteristics illustrate the operation of a tube under load provided that the emf of the direct

voltage source and the load in the output circuit of the tube are held constant.

When shown graphically, the above characteristics are called the *characteristic curves* (sometimes called simply characteristics or curves). A set of curves drawn for different values of the quantity chosen as parameter (independent variable) is called a *family of characteristics.*

The most important are static characteristics which show the relationship between an electrode current and an electrode voltage at no-load. These characteristics are specified in tube manuals.

In addition to static characteristics, use is made of response curves. An example is the frequency response curve which relates the amplification factor of a tube to the applied signal frequency.

In more detail, tube ratings and characteristics will be discussed in later chapters describing particular electron tubes.

1-3. ELECTRON EMISSION

Work Function. As is known from physics, the electric forces of interaction between an electron and a solid prevent the electron from leaving the solid. The Fermi energy, E_F, that is, the maximum energy which an electron in a metal has at $T = 0$ K, is not sufficient to overcome these forces.

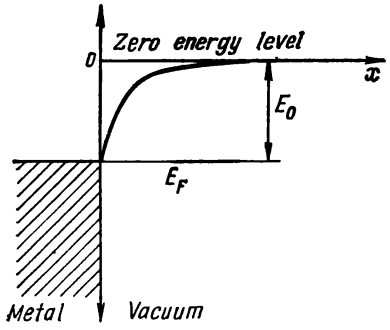


Fig. 1-4. Energy barrier near the metal surface

Let us take as a zero energy level the energy of an electron in a vacuum, infinitely distant from the surface of a solid and being affected by no forces. Then, the energy levels of the electrons in the solid should be below this zero level. Let their energy be negative, because a solid, as any other stable system, stores some energy determining the stability of the system. Accordingly, we may say that the energy level of an electron in a vacuum exceeds the Fermi level, E_F, by some value, E_0 (Fig. 1-4). This value equal to the difference between the energy levels in a vacuum and within the solid is called the *work function of the solid.*

The physical nature of the forces preventing the escape of an electron from a metal and determining the work that must be done by the electron in overcoming them is rather complex. The electron which has escaped from the metal is acted upon by the retarding field of a dipole layer formed by the electrons leaving the metal surface and the positive charges (called image charges) due to the

escaping electrons. Past the dipole layer, the electron is pulled back by a force equal to the force of attraction between the electron and the image charge induced within the metal.

The curve in Fig. 1-4 shows the work which should be done to move the electron for the distance x from the metal surface. In other words, the $E(x)$ curve depicts the energy barrier that exists at the interface between the metal surface and vacuum. In some cases, such as with the electric fields in the interelectrode space of an electron tube, it is more convenient to speak of a potential barrier. Then the potential $V = E/e$ is plotted as ordinate.

The work function, E_0, of metals is equal to several electron volts, being, for example, 1.88 eV for cesium, 2.4 eV for barium, and 4.5 to 4.6 eV for tungsten.

The work function can markedly be changed if the surface of the base metal is given a thin coat of some other material. Reduction in the work function results when the base metal is coated with materials electrically positive with respect to the base. In this case, the atoms of the coating material give up their electrons to the base metal. Because of this, ions are formed at the surface, which together with the negative charge of the base metal make up the dipole layer. The electric field of this layer accelerates the escaping electrons, thereby reducing the work function. The work function depends not only on the physical properties of the base metal and activating material, but also on the thickness of the activator. Reduction in the work function is most pronounced when the base metal is given a coat of an activator one atom thick. In this case, the work function for bariated tungsten is 1.56 eV and for thoriated tungsten, 2.63 eV.

Forms of Electron Emission. To free itself from the binding forces of a material, an electron has to do some work. This work cannot be done unless the electron receives some additional energy.

The ejection of electrons from the surface of a material is called *electron emission*. Electron emission from a solid into a vacuum can occur in one of four ways, as follows: (1) thermionic emission, (2) photoelectric emission, (3) secondary emission, and (4) field emission. As is seen, the forms of emission differ in the source of energy. The number of electrons leaving the solid and their speed in a vacuum depend on the amount of energy they acquire and also on the properties of the emitting body, that is, the cathode.

Thermionic Emission. When the energy is supplied thermally, by heating the solid, thermionic emission takes place. This has found the widest application in electron tubes.

Figure 1-5 shows the energy distribution function of electrons in a metal, dN/dE, which obeys the Fermi-Dirac statistics. At $T = 0$ K, the highest energy of an electron in a metal corresponds to the Fermi level, E_F. When the temperature of the metal increases, the fastest electrons may move to higher vacant energy levels. At

$T > 0$ K, the distribution function dN/dE is different: the probability that the energy levels above E_F will be occupied is no longer zero. At a certain temperature, the energy of the electrons becomes enough for them to escape from the metal, leave its surface, and move in

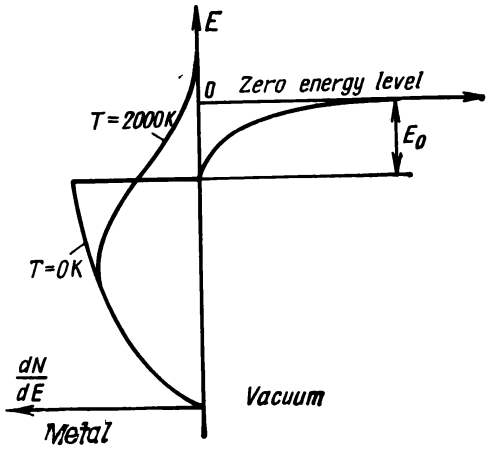


Fig. 1-5. Energy barrier near the metal surface and the distribution function dN/dE for the metal

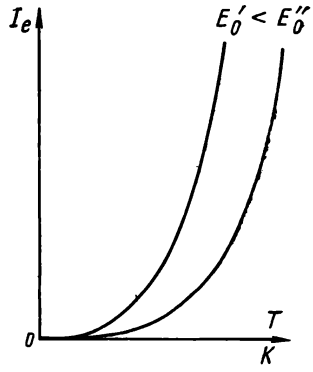


Fig. 1-6. Emission current as a function of temperature

a vacuum with a kinetic energy determined by the difference between their total energy and E_0.

The density current for thermionic emission, that is, the current per unit area of cathode surface is given by

$$j_e = A_0 T^2 \exp(-E_0/kT) \tag{1-1}$$

where

$$A_0 = 4\pi emk^2/h^3 = 120 \ A/(\text{cm} \times \text{K})^2 \tag{1-2}$$

m = electron mass
e = electron charge
k = Boltzmann's constant, $1{,}38 \times 10^{-23}$ J/K
h = Planck's constant, 6.63×10^{-34} J s.

Experiments have shown that A_0 may range from 10 to 300 for various materials.

As is seen from Eq. (1-1), the emission current density depends on the work function and temperature. Figure 1-6 shows the emission current, I_e, as a function of temperature for two cathodes having the same surface area, s ($I_e = j_e s$), but differing in the work function.

As already noted, thermionic emission is the most widely used form of emission. Thermionic cathodes have found application in all types of vacuum tubes, electron devices operating in the microwave region, electron-beam devices and in most gas-filled tubes.

The Schottky Effect. We have discussed thermionic emission on the assumption that no external electric field exists in a vacuum. But this is not exactly the case, because in most electron tubes the cathode is near the other electrodes whose potentials, in the general case, are other than zero. Accordingly, electrons leaving the cathode surface find themselves in an external electric field. The energy

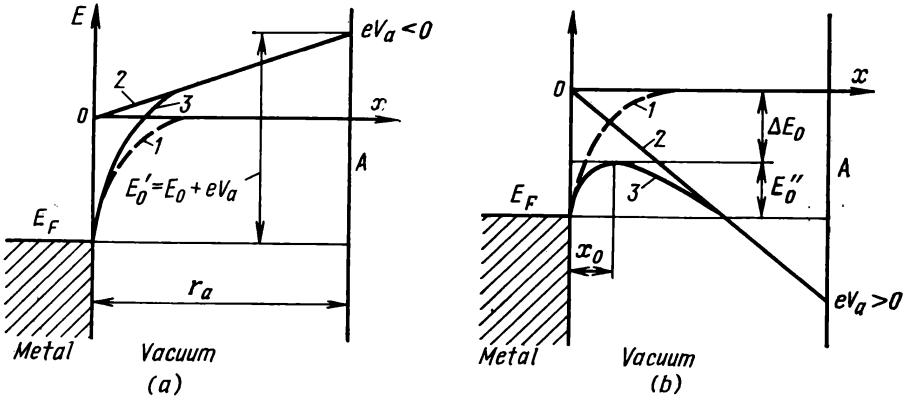


Fig. 1-7. Change in the energy barrier near the metal surface in the presence of an external electric field
(*a*) with a retarding field; (*b*) with an accelerating field

barrier near the cathode surface and, as a consequence, the conditions for electron emission are, therefore, different.

Figure 1-7*a* shows a change in the energy barrier near the cathode surface in the presence of a retarding external electric field. The resultant curve, *3*, is derived by adding together curve *1* which represents the energy barrier in the absence of an external field, and straight line *2* which represents variations in the potential between the cathode and electrode A whose potential is negative ($V_a < 0$). In moving away from the cathode for a distance of $x = r_a$, an electron having an energy of E_F has to do work equal to $E_0' = E_0 + e \mid V_a \mid$. In other words, it can reach electrode A only if it has received an additional energy, $E_0' > E_0$.

If electrode A is made positive ($V_a > 0$), the energy barrier will take the shape shown in Fig. 1-7*b*. Here, the height of the barrier is lower by ΔE_0 as compared with the case when there is no external field: $E_0'' = E_0 - \Delta E_0$ (at $x = x_0$).

Taking into consideration the effect of the external field, we may use for the emission current density an equation given in reference [1]

$$j_e' = j_e \exp(0.45\sqrt{\mathscr{E}}/T) \qquad (1\text{-}3)$$

where j_e is defined by Eq. (1-1) and $\mathscr{E}$ is the external field strength.

The increase in emission current owing to a reduction in the energy barrier (in the work function) under the action of an (accelerating) external field is commonly referred to as the *Schottky effect.*

Field Emission. Let us make electrode A still more positive. Figure 1-7*b* will still apply to the situation except that the height of the potential barrier E_0'' will be reduced, the width of the barrier will be very narrow, and it will approach the cathode. Due to the high field strength, $\mathscr{E}$, the potential barrier becomes so narrow that emission occurs by diffusion through the barrier. This is known as the "tunnel" effect. Now electrons are torn out, as it were, from the cathode surface by the strong electric field. Such emission is called *field emission.* Calculations show that field emission occurs at an external field strength $\mathscr{E}_{cr}$ of about 10^8 V/cm. However, experimental observations give $\mathscr{E}_{cr}$ = (approx.) 10^6 V/cm. This difference can mainly be ascribed to the roughness of the cathode surface.

Field emission is utilized only in some gas-filled devices (such as mercury-pool rectifiers and ignitrons), although the effect of an external electric field must also be taken into account when considering other types of emission utilized in many electron tubes.

Secondary Electron Emission. Secondary emission refers to the emission of electrons occurring as a result of other electrons striking a solid. The solid being bombarded is usually called a *secondary-electron emitter* or a *target.* When electrons (called primary ones) strike the surface of a target, some are resiliently reflected and the others penetrate into the solid and dissipate their energy by interacting with the crystal lattice and electrons of the target. As a result, a primary electron may either entirely dissipate its energy and remain within the solid or spend only some of its energy, change the direction of its travel and escape from this solid. The latter case corresponds to the nonresilient reflection of primary electrons. The energy dissipated by a primary electron within the solid may be converted into the kinetic energy of its electrons, and these will, as a consequence, escape from the target. These are called *secondary electrons proper.*

Thus, the bombardment of a target by primary electrons results in a reverse stream of secondary electrons, made up of not only the secondary electrons proper, but also resiliently and nonresiliently reflected electrons. The relative numbers of the three types of electrons and the total number of secondary electrons depend on the energy of primaries, physical and chemical properties of the solid, cleanness of its surface, angle of incidence of primary electrons, and some other factors.

Quantitatively, secondary emission is described in terms of the ratio of the total number of secondary electrons, n_2, to the number of primary electrons, n_1. This ratio is known as the *secondary-elec-*

tron yield or the *secondary-emission ratio*:

$$\sigma = n_2/n_1 \qquad (1\text{-}4)$$

Figure 1-8 shows the secondary-emission ratio versus primary-electron energy E_1. As the energy E_1 rises, the number of secondary electrons increases until the secondary-emission ratio reaches its maximum at a certain value E_{1cr}; E_{1cr} and σ_{max} are different for different materials.

As the energy of primary electrons, E_1, increases, they bore deeper into the target. As a result, the deeper-seated secondary electrons transfer the imparted energy to still other electrons and are slowed down before reaching the surface. The number of resiliently reflected electrons decreases fo the same reason. Thus, the secondary-emission ratio goes down.

For most metals and semiconductors, σ_{max} is slightly above unity. Complex compounds containing materials with a low work function, for example, cesium, have a σ_{max} of several unities.

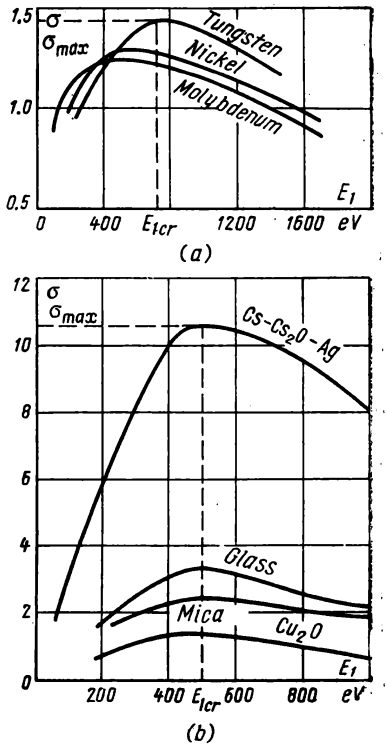


Fig. 1-8. Secondary-emission ratio as a function of the energy of primary electrons
(*a*) for metals; (*b*) for semiconductors and dielectrics

1-4. THERMIONIC CATHODES

Characteristics and Ratings. Most electron tubes use thermionic, or hot, cathodes which supply electrons by thermionic emission. The performance of a cathode is expressed in terms of several variables. One of them is *emission current density*, that is, the emission current per unit area of the cathode surface, defined by Eq. (1-1)

$$j_e = A_0 T^2 \exp(-E_0/kT)$$

As is seen, the value of j_e characterizing the emissive capacity of a cathode depends on its physical properties (coefficient A_0), temperature T and work function E_0.

Another variable is the *operating temperature of a cathode*, T_{op}, at which the cathode shows the best performance. T_{op} is chosen from considerations of energy balance, including the thermal energy received by the cathode, bombardment of the cathode by electrons

and ions, loss of energy by radiation, heating of the cathode holders, evaporation of the cathode material, electron emission and so on.

A very important variable, showing how efficient a cathode is, is the ratio of emission current to filament dissipation

$$H = I_e/V_f I_f = I_e/P_f \tag{1.5}$$

Here, I_f and V_f are the filament current and voltage, respectively; $I_e = j_e s$ is the emission current. As a rule, this ratio, called the *emission efficiency*, is expressed in milliamperes per watt (mA/W).

Experimental observations have shown that about 70% of the applied power is dissipated by the filament as radiation and only an insignificant part is utilized in electron emission.

The cathode variables also include the *life of a cathode*, which is determined on the basis of its emission. A cathode is said to "have burned out", if its emission current has dropped by twenty per cent of its nominal value.

The causes of cathode failure are diverse and depend on the type of cathode and its operating conditions. One of the principal causes is the sputtering of the cathode material when it is heated. Cathode emission can also be reduced because the activator layer is destroyed by ion bombardment, oxides or other chemical compounds increasing the work function are formed on its surface, etc.

Construction of Thermionic Cathodes. There are two groups of cathodes, namely *directly heated* (also known as *filament* or *filamentary*) *cathodes* and *indirectly heated* (also called *heater*) *cathodes*. In the former case, the filament current I_f flows directly through a cathode. These cathodes are usually in the form of a thin wire or ribbon bent to the shape of the other electrodes generally flat or cylindrical. Cathodes are carried by current-conducting supports. Some of the most commonly used forms of directly heated cathodes appear in Fig. 1-9*a* and *b*.

Heater cathodes are commonly made in the form of a small metal cylinder that fits over a thin wire called the heater. The cylinder is given a coat of an activator from the outside and is slipped over the heater after the latter has been coated with an insulating material that will permit operation at high temperatures. The filament current flows in the heater and heats it, and the heat is then supplied to the cathode. Some of the most widely used forms of indirectly heated cathodes are shown in Fig. 1-9*d*.

A major advantage of indirectly heated over directly heated cathodes is that they can be supplied with alternating current. Owing to this, filament voltage sources can be made much simpler.

Directly heated cathodes usually operate on direct current, because if, for example, a cathode were heated with a current at 50 Hz, the anode (plate) current would contain the ripple at 50 or 100 Hz. In audio circuits, this interference would appear as hum.

Operation of Thermionic Cathodes. The performance of a cathode is usually described in terms of its filament current I_f and filament voltage V_f whose nominal values correspond to the operating temperature of the cathode. The relationship between these variables, $I_f=f(V_f)$, is called the *filament characteristic of a cathode* (Fig. 1-10).

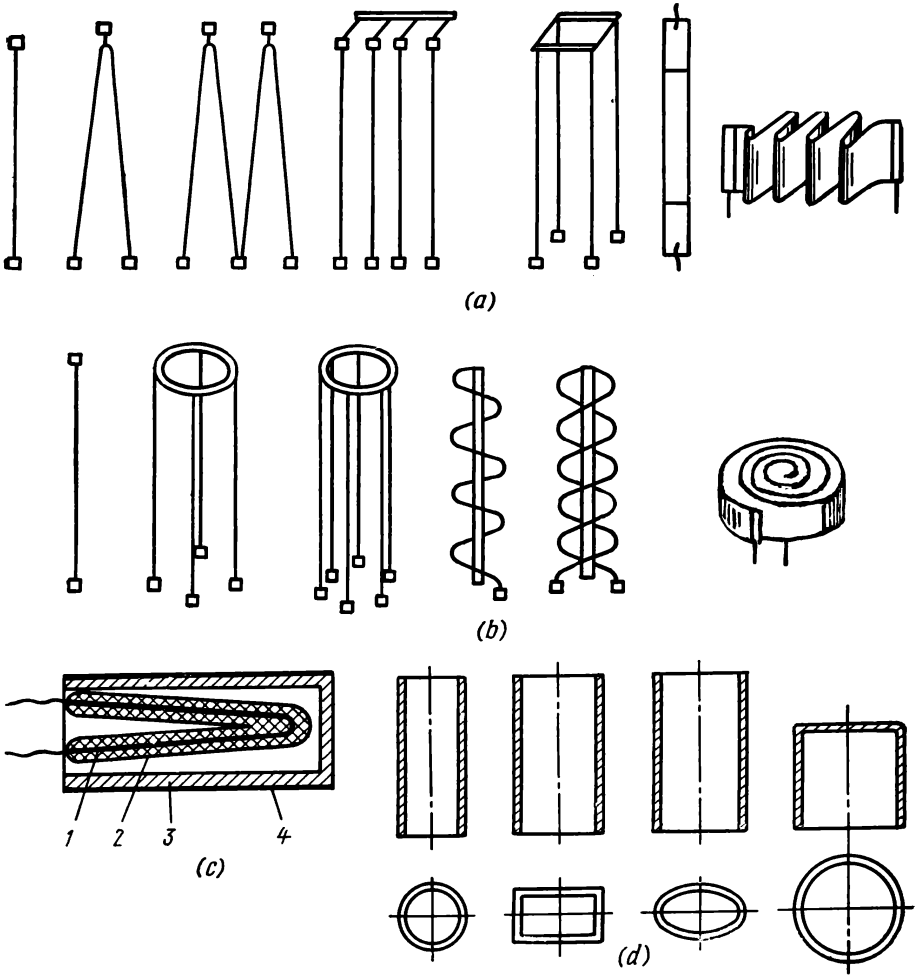


Fig. 1-9. Forms of cathodes

(*a*) directly heated cathodes for flat electrodes; (*b*) directly heated cathodes for cylindrical electrodes; (*c*) construction of a heater cathode; *1*—heater; *2*—insulating coating; *3*—metal cylinder; *4*—activator layer; (*d*) different shapes of heater cathodes

It is nonlinear, because the resistance of the filament rises with temperature. So, for some cathodes (for example, a tungsten cathode) it is advisable to raise the filament current gradually and reach the nominal value of I_f when the cathode has attained its operating temperature.

The temperature of a cathode can be maintained by holding constant either the filament current or the filament voltage which can be monitored by an ammeter or a voltmeter, respectively. The former technique is seldom used, because, with time, a directly heated cathode or the heater of an indirectly heated cathode grows thinner, its resistance increases, and its mass is reduced. If the filament current were held constant, the life of the cathode would be cut down because, towards the end of its service life, it would operate with overheating.

Fig. 1-10. Filament characteristic of a cathode

So, more frequently, the performance of a cathode is maintained by holding constant the filament voltage. In this case, the life of the cathode is doubled or trebled as compared with the previous technique, although towards the end of its life it operates with underheating and, as a consequence, its efficiency is reduced.

Types of Thermionic Cathodes. The cathodes or emitters as used in present-day thermionic electron tubes may be divided into three groups, namely *pure metal emitters*, *atomic-film emitters*, and *oxide coated emitters*.

Pure metal emitters should have a low work function and high melting temperature because, as follows from Eqs. (1-1) and (1-5), a reduced work function and increased melting temperature result in a higher emission efficiency. Metals used in pure metal cathodes should be viscous, because the cathodes are made in the form of a thin wire. They should also be easily degassable and stand up well to oxidation. The most commonly used are tungsten emitters.

Tungsten cathodes offer stable emission current and stand up well to ion bombardment. Because of this, they are especially good for high-power transmitting tubes and electrometer tubes.

Present-day *atomic-film emitters* include thoriated-tungsten filaments, bariated-tungsten emitters (which in fabricated form are known as *L* cathodes), matrix emitters, and so on.

Figure 1-11 shows the construction of a dispenser-type *L* cathode with a flat and a cylindrical emitting surface. The internal cavity of a molybdenum cylinder contains a heater. The external reservoir is filled with an active material: barium carbonate — strontium carbonate, thoria, etc. On the outside, the active material is separated from a vacuum by a plate or cylinder of porous tungsten or molybdenum welded to the first cylinder. As the *L* cathode is gradually heated, the active material decomposes, gases are removed from the envelope, and the oxides are reduced by reaction with tungsten. Barium atoms diffuse to the surface of the porous tungsten body and

form an active layer. *L* cathodes stand up well to ion bombardment and gas poisoning (after it ceases emitting electrons due to the action of gases, the *L* cathode recovers its emission activity very quickly). Dispenser-type *L* cathodes are used in electron-beam devices and electron tubes operating in the microwave region.

The impregnated type of *L* cathode is simpler in construction. It has no cavity for the active material; instead it is dispersed through-

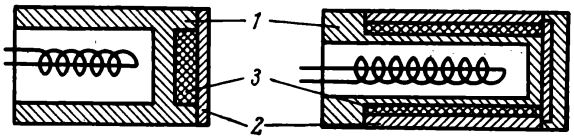


Fig. 1-11. Construction of the L-cathode
1—molybdenum cylinder; *2*—porous tungsten body; *3*—activator

out the porous tungsten body. As compared with the dispenser *L* cathode, the impregnated *L* cathode is heated more evenly and has a greater efficiency.

Matrix cathodes are a modification of the *L* cathode. They are fabricated by sintering or pressing powdered metal oxides (nickel, tungsten and iron oxide) and alkali-earth metal carbonates. Their emission efficiency is almost the same as that of *L* cathodes.

The more recent additions to the range of atomic-film emitters have been boride cathodes. The active material in them is lanthanum hexaboride or barium hexaboride. Boride cathodes are made either by giving a molybdenum or tantalum base layer a coat of active material, or by pressing cylinders or discs which are then installed in mountings of the required shape. Boride cathodes have high emission capacity and show good performance in the presence of strong external fields.

Oxide-coated cathodes are the most commonly used in various vacuum and gas-filled tubes. The base metal is nickel or tungsten. The oxide coating usually consists of the mixture of oxides of the alkali-earth metals barium, strontium, and calcium. The best emission capacity is offered by oxide-coated cathodes using nickel with addition of tungsten.

As the barium is reduced, free barium atoms are formed in the vertices of the crystal lattice. In this case, the barium acts as a donor impurity and the oxide coat becomes an *N*-type semiconductor. Some free barium atoms diffuse to the surface to make a monatomic layer. The barium atoms vaporizing from the coating are replenished by those coming from the inside of the deposit.

A major advantage of oxide-coated cathodes is that they can show high emission in pulse work.

Along with high efficiency and current density (especially in pulse work), oxide-coated cathodes have a few drawbacks. The barium vaporizing from the cathode surface contaminates other electrodes. Because of this, their secondary-emission ratio rises to a point where it may disturb the operation of the tube. Besides, oxide-coated cathodes are considerably affected by ion bombardment, so, they cannot be used in tubes operating at high voltages where the speed of bombarding ions is great. The oxide coating is not uniform, so spots with high resistance and strong electric field may appear on its surface. When a very heavy anode current is used, these spots are heavily heated and the coating may break down (eject sparks).

In order to remove these disadvantages, several modifications of oxide-coated cathodes have been designed. In one of them, known as the oxide-grid cathode, the base metal gives support to a metal grid whose meshes are filled with an oxide. This improves the conductivity of the cathode and, as a consequence, reduces sparking.

Good performance is shown by barium-oxide and thorium-oxide cathodes. They stand up well to ion bombardment and gas poisoning (black-outs), recover their emission activity after black-out periods very quickly and permit high emission current densities in continuous duty.

1-5. ELECTRIC CURRENT IN A VACUUM

Induced Current. Let us imagine two flat electrodes, a cathode, K, and an anode, A, located in a vacuum and spaced some distance r_a apart (Fig. 1-12). These electrodes are connected by an external conductor and grounded. Each of them has a surface area of Q.

If we place a negative electric charge ($-q$) at a distance x from electrode K, positive charges q_K and q_A will be induced on both electrodes. These charges are inversely proportional to the distance from the charge $-q$ to the respective electrodes

$$q_A/q_K = x/(r_a - x) \tag{1-6}$$

As is known from electrostatics, the sum of all charges in a system is zero

$$-q + q_K + q_A = 0 \tag{1-7}$$

Simultaneous solution of these equations gives

$$q_A = qx/r_a \tag{1-8}$$

$$q_K = q\,(1 - x/r_a) \tag{1-9}$$

If we move the charge $-q$ from electrode K to electrode A, the charges induced on both electrodes will vary with x: q_K will decrease and q_A will increase. Because the induced charges are re-distribu-

ted, an equalizing current, known as *induced current*, will flow in the external circuit

$$i_{ind} = dq_A/dt = d(qx/r_a)/dt = (q/r_a)/(dx/dt) \quad (1\text{-}10)$$

Here, $dx/dt = v$ is the speed of charge motion. So

$$i_{ind} = qv/r_a \quad (1\text{-}11)$$

The direction of the induced current in the external circuit coincides with the direction of motion of positive charges, so, in our case, the current in the external circuit flows from electrode K to electrode A.

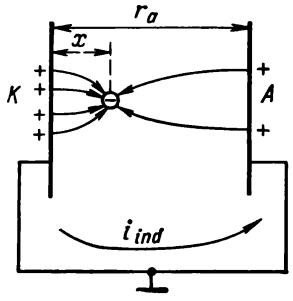


Fig. 1-12. Generation of induced charges

Composite Current in a Diode. In electron devices, the potential at the cathode is taken to be zero; the potentials at the other electrode(s) are usually non-zero.

Suppose that we have applied to the anode (electrode A) a positive potential V_A with respect to the cathode. Let us also assume that the cathode has no thermal lag, that is, it requires no warm-up period after the filament battery has been turned on. As soon as the filament is on, the cathode surface begins to emit electrons, and these move to the anode under the action of its positive field. As soon as the first electrons leave the cathode, positive charges are induced at the cathode and anode, and an induced current begins to flow in the external circuit. As is seen, a current starts flowing in the external circuit the same instant as the first electrons are emitted. During the time it takes the leaving electrons to reach the anode, the entire circuit may be divided into three regions.

As soon as the cathode begins to emit electrons, an induced current appears in the external circuit, whose instantaneous value can be derived from Eq. (1-11). It should be remembered that now q represents the charge due to all electrons present in the volume $V = Qx$ at a given time.

In the space between the cathode and the front of the first electrons, there is a flow of transfer current due to the motion of electric charges. Its density is proportional to the space-charge density ρ and electron velocity v

$$j_{tr} = \rho v \quad (1\text{-}12)$$

As the electrons move from plane x to the anode, the transfer current does not change in value, because the cathode emission is independent of time, and at all subsequent moments the same number of electrons travelling at a constant speed passes through plane x.

In the third region, that is, between plane x and the anode which the electrons have not reached yet, the circuit is completed by a dis-

placement current whose density is determined by a change in electric displacement $D = \varepsilon_0 \mathscr{E}$

$$j_d = \partial D/\partial t = \varepsilon_0 \partial \mathscr{E}/\partial t \qquad (1\text{-}13)$$

Initially, when the cathode has not yet begun to emit electrons, the current in the circuit is zero and the strength of the electric field in the vacuum is determined by the potential difference V_A and the distance between the electrodes, r_a. After the first electrons have left the cathode, the electric field due to the potential difference V_A is added together with the fields due to the electrons and induced charges. As the first electrons travel from the cathode to the anode, the total negative charge in the interelectrode space increases and the rate of change of the field strength decreases. The displacement current falls, too. When the first electrons reach the anode, the circuit attains a steady state. Here, in contrast to the transient process discussed just now, the total electric charge in the vacuum space remains unchanged, and so does the electric field strength; the displacement current is equal to zero. The induced current in the external circuit does not change, either, (q = const) and is equal to the transfer current. This is so because the total charge in an elementary layer dx (Fig. 1-13) is

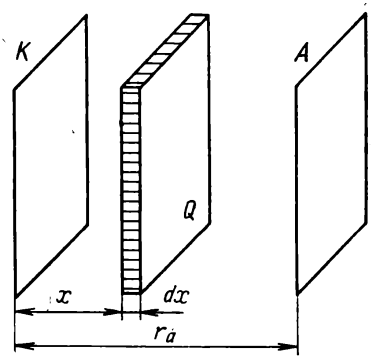


Fig. 1-13. Explaining the derivation of an equation for current in a tube

$$q_x = Q\rho dx \qquad (1\text{-}14)$$

an the total charge in the entire volume between K and A is

$$q = \int_0^{r_a} Q\rho\, dx \qquad (1\text{-}15)$$

Using Eq. (1-11), we may write

$$I_{ind} = \int_0^{r_a} (Q\rho v/r_a)\, dx \qquad (1\text{-}16)$$

Using Eq. (1-12) and integrating, we get

$$I_{ind} = \int_0^{r_a} (j_{tr}Q/r_a)\, dx = j_{tr}Q = I_{tr} \qquad (1\text{-}17)$$

As is seen, in the steady state, the current induced by travelling electrons in the external circuit is equal to the transfer current in the vacuum.

Chapter Two

VACUUM DIODES

2-1. PHYSICAL PROCESSES IN THE VACUUM DIODE

Definition. Construction. A vacuum tube with only two electrodes, namely a cathode and a plate (or anode), is called a *vacuum diode.* Here, the thermionic cathode operates as an emitter, that is, it emits electrons, and the plate serves as a collector of electrons. According

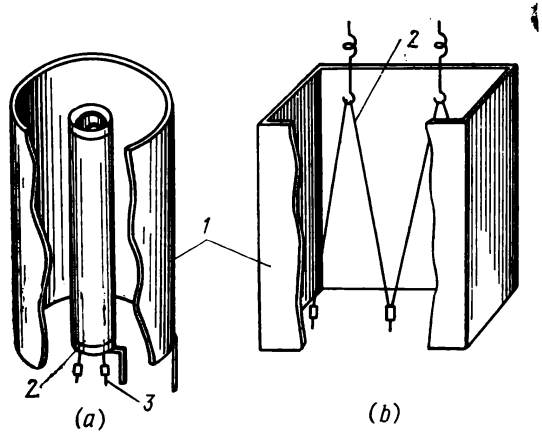


Fig. 2-1. Construction of a vacuum diode
a) with cylindrical electrodes and indirectly heated cathode; (*b*) with flat electrodes and directly heated cathode, *1*—plate; *2*—cathode; *3*—heater

to the shape of the cathode, the plate is made in the form of a cylinder (Fig. 2-1*a*) or a box (Fig. 2-1*b*). Both electrodes are placed in an evacuated glass, metal, or metal-ceramic envelope.

Principle of Operation. As with all vacuum tubes, it is customary to take the cathode potential in a diode to be equal to zero ($V_k = 0$); the potentials at all other electrodes are determined with respect to this value.

When filament current I_f is applied to heat the cathode, thermionic emission takes place. The electrons escaping from the cathode

travel towards the plate, because it is held at a positive voltage, V_p. This brings about a plate current I_p directed from the cathode to the plate in the external circuit. If the plate were negative, the electrons leaving the cathode would find themselves in a retarding field and return to the cathode, so the current in the external circuit would be zero. As is seen, current in the diode can flow only in one direction, namely from the plate to the cathode, when the plate is at a potential which is higher than that of the cathode. This is the main distinction of a diode. Devices conducting current only in one direction are sometimes called electric valves*.

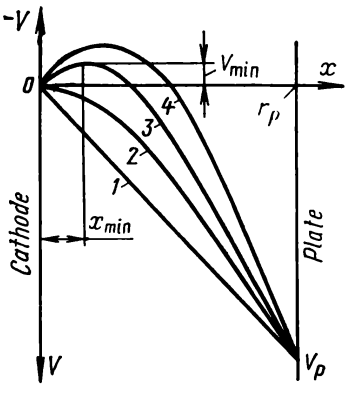


Fig. 2-2. Potential distribution in a vacuum diode at $V_p = \text{const}$

The value of plate current determined by the number of electrons reaching the plate depends on both emission current I_e and plate voltage V_p. As the emission current rises, the number of electrons leaving the cathode and travelling towards the plate increases, and so does the plate current. The plate voltage also affects the number of electrons reaching the plate. When the plate is held at a low positive potential, not all electrons reach it; those having low initial velocities return to the cathode. Now, let us consider these phenomena in more detail.

Potential Distribution in the Diode. Let the plate and the cathode be unbounded planes such as shown in Fig. 2-2. Here, the distances from the cathode are laid of as abscissas, and the positive potentials downwards as ordinates. We shall examine the potential distribution in the cathode-plate space at different filament voltages, V_f, and a fixed plate voltage, V_p.

If $V_f = 0$ and no emission takes place, the diode may be regarded as a parallel-plate capacitor whose potential difference is V_p. In this case, the potential distribution in the cathode-plate space is represented by straight line *1* (Fig. 2-2) which joins the points corresponding to the cathode potential $V_k = 0$ and the plate potential V_p. When the filament voltage rises, the electrons leaving the cathode gang up in the interelectrode space as a cloud called a *space charge*. This charge alters the potential distribution (curve *2* in Fig. 2-2). Since the electrons making up the space charge are negative, the potential in the cathode-plate space goes down, though all points

* Most commonly, they are used to rectify alternating current. So, by extension, they are more usually called rectifier tubes. — *Translator's note.*

remain at positive potential. At any point on curve *2*, the vector of the electric field strength is directed from the plate to the cathode. So in an accelerating field, all electrons escaping from the cathode make for the plate. In this case, plate current is equal to emission current. The condition in which all of the cathode emission is used up is called *saturation*.

As the filament voltage becomes higher, emission increases, and so does the space charge. Electrons having low initial velocities are driven back to the cathode by the negative space charge due to other electrons. The density of the electron cloud near the cathode increases to a point where it forms a negative potential region (curve *3* in Fig. 2-2) whose minimum, $V_{\min}$, is usually within a few hundredths or tenths of a millimetre of the cathode surface. Thus, there is a retarding electric field near the cathode ($0 < x < x_{\min}$); its strength vector is directed from the cathode to the plate. To overcome this field, the initial velocity v_0 of electrons leaving the cathode should exceed a certain value determined by $V_{\min}$

$$v_0 > \sqrt{2\,(e/m)\,V_{min}} \qquad (2\text{-}1)$$

If the initial velocity of an electron is below this value, the electron will not be able to overcome the potential barrier. It will slow down to a stop, and the field will push it back to the cathode. Accordingly, the retarding field region (from 0 to $x_{\min}$) contains not only electrons travelling towards the plate, but also electrons falling back to the cathode. At a constant filament voltage, a dynamic equilibrium sets in, so that the number of electrons reaching the plate and falling back to the cathode is equal to the number of electrons emitted by the cathode. Therefore, plate current is smaller than emission current.

As the cathode becomes still hotter, a dynamic equilibrium corresponding to a higher emission current sets in. However, in this case, plate current grows not so fast as emission current because the potential barrier near the cathode becomes higher (curve *4* in Fig. 2-2).

Curves *3* and *4* represent what is known as the *space-charge-limited operation of the diode*. In this case, in contrast to saturation, not all electrons emitted by the cathode reach the plate, that is, plate current is smaller than emission current.

Figure 2-3 shows the potential distribution in a vacuum diode for several values of plate voltage and a constant filament voltage. If the cathode is hot and the plate voltage is zero ($V_{p1} = 0$), the electrons emitted by the cathode make up a negative space charge in the interelectrode space. Here, the potential distribution is represented by curve *1*. At a low plate voltage (curve *2* in Fig. 2-3), the negative potential region remains near the cathode. This potential barrier prevents slow electrons from travelling to the plate. Electrons with higher energies overcome it, reach the plate and establish plate cur-

rent which is smaller than emission current. At a higher plate voltage ($V_{p3} > V_{p2}$), the density of the space charge near the cathode decreases, the potential barrier is reduced, and its hump is now closer to the cathode (curve *3* in Fig. 2-3). The operation of the diode continues to be space-charge-limited. At a certain value of plate voltage, V_{p4}, the tube is running at saturation (curve *4* in Fig. 2-3). In this case, all points of the cathode-plate space are at positive potential, the field strength vector is directed from the plate to the cathode, and all electrons emitted by the cathode reach the plate.

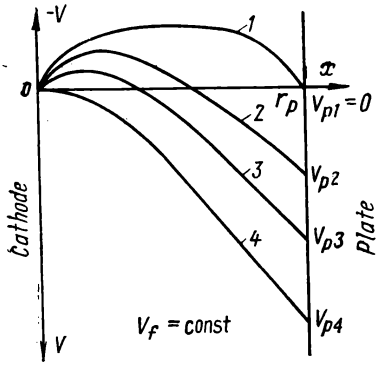


Fig. 2-3. Potential distribution in a vacuum diode at $V_f = \text{const}$

Space-charge-limited operation with a negative potential region near the cathode is the main type of operation for a vacuum diode, and, indeed, for all vacuum tubes. As a rule, all electron tubes operate at a constant (nominal) filament voltage and, as a consequence, at a constant emission current. It does not pay to control plate current by reducing or increasing the filament. Even very small cathodes have thermal lag, so with this method of plate current control, electron tubes would lose one of their most valuable properties — freedom from time lag. In diodes, control is achieved by varying plate voltage; in other electron tubes plate current is controlled by grids.

As is seen from Fig. 2-3, plate voltage affects the value and position of the potential barrier in the negative potential region near the cathode. When plate voltage increases, the barrier decreases and shifts towards the cathode, and the density of the space charge near the cathode goes down due to a rise in the number of electrons overcoming the barrier and travelling to the plate. In contrast, a fall in plate voltage brings about a rise in the space charge density owing to an increase in the number of electrons returning to the cathode.

Thus, the space charge serves as a reservoir, or a source, supplying electrons to the plate. It is continuously replenished by electrons emitted by the cathode, so that it is in a dynamic equilibrium which, naturally, depends on plate voltage.

2-2. PLATE CURRENT AS A FUNCTION OF PLATE VOLTAGE

The Three-Halves Power Law. Let us determine the relationship between plate current and plate voltage in a vacuum diode having unbounded plane-parallel electrodes (Fig. 2-4). In this case, we

may neglect the edge effect and consider the field between the plate and the cathode to be uniform. Assume also that near the cathode surface ($x = 0$), $V_k = 0$, $\partial V/\partial x = -\mathscr{E}_x = 0$, electron initial velocity is $v_0 = 0$ and plate potential is V_p.

At $v_0 = 0$, the velocity of an electron at any point of the cathode-plate space with potential V is

$$v = \sqrt{2\,(e/m)\,V} \tag{2-2}$$

Taking into consideration that there is a space charge in the cathode-plate space, we use Poisson's equation which, at $\mathscr{E}_y = \mathscr{E}_x = 0$

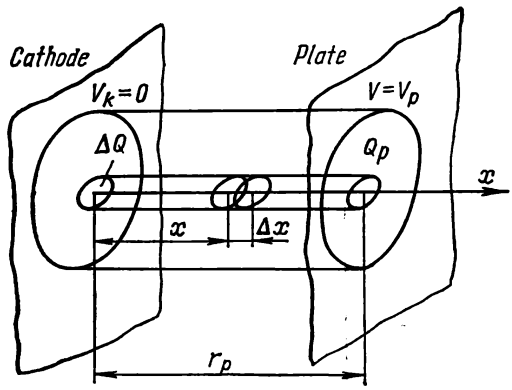


Fig. 2-4. Explaining the derivation of the three-halves power law

and under conditions of a vacuum, may be written

$$\partial^2 V/\partial x^2 = -\rho/\varepsilon_0 \tag{2-3}$$

Here, ρ is the charge density, C/m^3, and ε_0 is the permittivity of free space, 8.85×10^{-12} F/m.

As is known from physics

$$\rho = -j/v \tag{2-4}$$

or, recalling Eq. (2-2), we may write

$$\rho = -j\sqrt{m/2e}/\sqrt{V} \tag{2-5}$$

Substituting Eq. (2-5) for ρ in Eq. (2-3), we re-write Poisson's equation as

$$\partial^2 V/\partial x^2 = (j/\varepsilon_0)\,(\sqrt{m/2e}/\sqrt{V}) \tag{2-6}$$

On multiplying both sides by $2\,\partial V/\partial x$ and integrating from 0 to x, we get

$$(\partial V/\partial x)^2 = (4j/\varepsilon_0)\sqrt{mV/2e} \tag{2-7}$$

The integration constants are equal to zero because, at $x = 0$, $V_k = 0$ and $dV/dx = 0$.

Extracting the root of both sides, separating the variables, and integrating Eq. (2-7), we get

$$4/3\sqrt[4]{V^3} = 2\sqrt{j/\varepsilon_0}\sqrt[4]{m/2e}\,x \tag{2-8}$$

Hence, the current density which is the same for any cross-section of the cathode-plate space may be written

$$j = (4\varepsilon_0/9)\sqrt{2e/m}\,\sqrt{V^3}/x^2 \tag{2-9}$$

Substituting $x = r_p$ and $V = V_p$, and also the numerical values for e, ε_0 and m in the equation gives the current density (in A/cm^2) at the plate surface

$$j_p = 2.33 \times 10^{-6} V_p^{3/2} r_p^{-2} \tag{2-10}$$

Here, V_p is in volts and r_p is in centimetres.

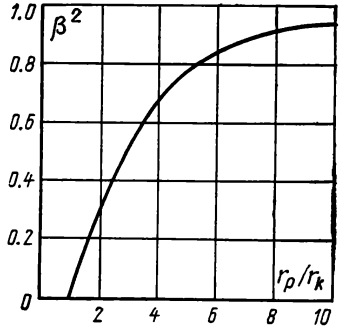


Fig. 2-5. Correction factor β^2 as a function of r_p/r_k

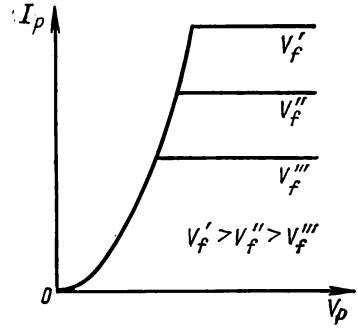


Fig. 2-6. Plate characteristics of an ideal diode

To find the value of plate current in amperes, we multiply Eq. (2-10) by Q_p, the *effective plate area*

$$I_p = 2.33 \times 10^{-6} Q_p / r_p^{-2} V_p^{3/2} \tag{2-11}$$

or

$$I_p = G V_p^{3/2} \tag{2-12}$$

where

$$G = 2.33 \times 10^{-6} Q_p / r_p^2 \tag{2-13}$$

is constant for a given type of electron tubes.

Equation (2-11) expresses the *three-halves power law** relating plate current to plate voltage, $I_p = f(V_p)$.

* This law is also known as the Child-Langmuir-Schottky equation, and the constant G is called the perveance of a tube.— *Translator's note.*

Equation (2-11) was derived for a planar diode. For cylindrical electrodes, r_p designates the radius of the plate. Besides, the denominator on the right-hand side of Eq. (2-11) should be multiplied by the correction factor β^2. Figure 2-5 shows β^2 as a function of r_p/r_k, where r_p and r_k are the radii of the plate and cathode, respectively. The quantity Q_p is called the effective plate area because it is the part of the plate surface that collects electrons. For cylindrical electrodes, it is equal to the internal surface area of the plate because the cathode is usually longer than the plate, and electrons strike all of the internal surface of the plate.

Characteristics of an Ideal Diode. The three-halves power law expressed by Eq. (2-11) relating plate current to plate voltage holds only for space-charge-limited operation. Under conditions of saturation, the plate current of an ideal diode does not depend on plate voltage and is determined only by filament voltage. The characteristics of an ideal diode are given in Fig. 2-6.

2-3. STATIC CHARACTERISTICS OF THE VACUUM DIODE

Figure 2-7 shows a family of characteristic curves representing plate current as a function of filament voltage, $I_p = f(V_f)$, with plate voltage held at a constant level for each characteristic. These curves

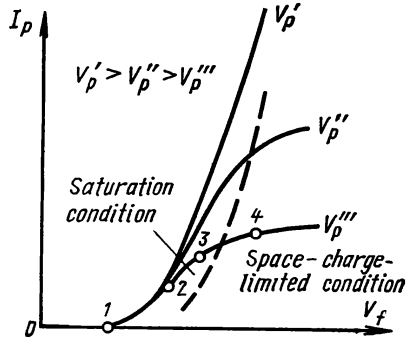


Fig. 2-7. Plate current as a function of filament voltage

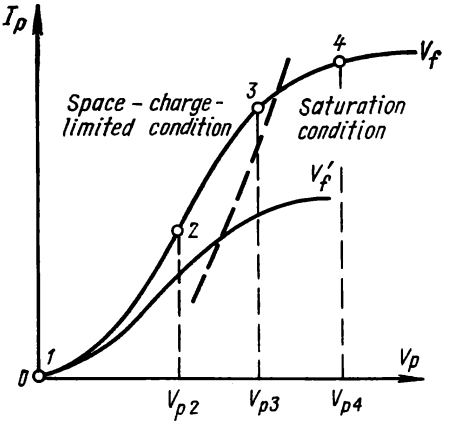


Fig. 2-8. Family of plate characteristics of a vacuum diode

have two distinct regions. Within the first region corresponding to low values of V_f, plate current increases with increasing filament voltage according to a law similar to that relating emission current to filament voltage or cathode temperature, Eq. (1-1). This ascending region is the emission characteristic, $I_e = \varphi(V_f)$, and corres-

ponds to saturation: all electrons emitted by the cathode reach the plate, $I_p = I_e$. Points *1* and *2* on the characteristic plotted at $V_p''' = = \text{const}$ correspond to curves *1* and *2* in Fig. 2-2, which represent the potential distribution in the cathode-plate space.

As V_f increases further, plate current rises more slowly since the diode is now operating under the space-charge-limited conditions (see curves *3* and *4* in Fig. 2-2). In this region, the curves no longer correspond to the emission characteristic, because $I_p < I_e$. An increase in plate voltage shifts the sloping part of the characteristic to the right because the space charge displays its limiting effect at high values of V_f.

Plate Characteristic. Figure 2-8 shows a family of plate characteristics of a vacuum diode using a tungsten cathode. One of the characteristics (plotted at $V_f = \text{const}$) shows points marked *1* through *4* and the values of plate voltage for which potential distribution curves are plotted in Fig. 2-3. Here, the steep regions correspond to the space-charge-limited operation, and the sloping ones to saturation. These characteristics considerably differ from those of an ideal diode (Fig. 2-6).

These differences are related to some physical processes in the vacuum diode that were not taken into consideration when the three-halves power law was derived. Let us see what causes the real characteristics to differ from the idealized ones.

Real characteristics do not always start from the origin of coordinates. For most tubes, the plate characteristic has a tail which implies that even at $V_p = 0$, there is a small plate current flowing in the tube. The plate current flowing at low negative voltages V_p is due to electrons having high initial velocities. In deriving Eq. (2-11), we assumed that their initial velocity was zero.

With directly-heated power diodes, the initial region of the characteristics is sometimes shifted to the right, so, at low positive plate voltages, plate current is zero. This can be accounted for by the action of a magnetic field established near the cathode by the filament current. This magnetic field tends to drive electrons escaping from the cathode to fall back.

The ascending region of a real characteristic also somewhat differs from the idealized characteristic. As a rule, real characteristics run more gently. This is mainly explained by the nonuniform potential and temperature distribution at the cathode surface. Indeed, in deriving the three-halves power law, we assumed that any point of the cathode surface was at constant zero potential, that is, the cathode was equipotential, and held at the same temperature. Under real conditions, however, neither of these assumptions is true. Owing to increased heat abstraction, the temperature at the ends of a cathode and at the electrode supports is lower than elsewhere; besides, directly heated cathodes are not equipotential.

The same factors affect the shape of that segment of the characteristic which corresponds to the transition from the space-charge-limited operation to saturation. As is seen from Fig. 2-8, in real characteristics, this transition is smooth. Saturation is reached earlier within the regions with a lower cathode temperature and a higher potential difference between the plate and the cathode. How smooth the transition is depends also on the spread in velocity between electrons travelling to the plate. For electrons having higher initial velocities, saturation takes place somewhat earlier.

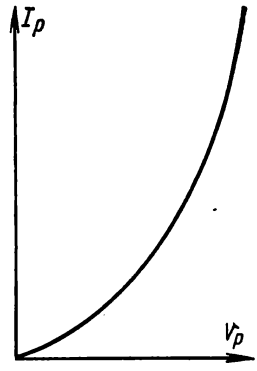


Fig. 2-9. Plate characteristic of a vacuum diode using an oxide-coated cathode

The difference from the idealized characteristic is most pronounced in the saturation region. As is shown in Fig. 2-8, the sloping part of the characteristic is typical only of tubes utilizing tungsten cathodes. No saturation occurs in vacuum tubes with an oxide-coated or any other cathode whose resistivity is fairly high (Fig. 2-9). This can be explained, firstly, by the Schottky effect (Sec. 1-3) and, secondly, by additional heating of the cathode due to the tube current.

2-4. THE LOAD LINE

In practical circuits, the external circuit of a diode contains a load. Let us discuss the simplest case, when the plate circuit of a diode contains a load resistor, R (Fig. 2-10*a*). According to Kirchhoff's voltage law, we may write for a loop carrying the current I_p

$$E_p = I_p R + V_p \tag{2-14}$$

where V_p is the voltage at the plate of a diode with respect to the cathode.

This equation is in two unknowns, I_p and V_p; their relationship, $I_p = f(V_p)$, is the plate characteristic of the tube. So, most conveniently, Eq. (2-14) can be solved graphically by drawing a plot on the plate characteristic.

Let us rewrite Eq. (2-14) as follows

$$I_p = E_p/R - V_p/R \tag{2-15}$$

Since $E_p/R = \text{const}$, this is the equation of a straight line plotted in coordinates where I_p is laid off along the y-axis and V_p along the x-axis. The *load line*, as this straight line is called, can easily be drawn on the plate characteristic of the diode. Assuming $I_p = 0$, we find the intercept made by this line on the voltage axis, equal to

E_p. On putting $V_p = 0$, we find the intercept made by the load line on the current axis, equal to E_p/R.

Instead of locating one point on each axis, the load line may be constructed by its slope, tan α = (m/n) $(1/R)$, where m and n are

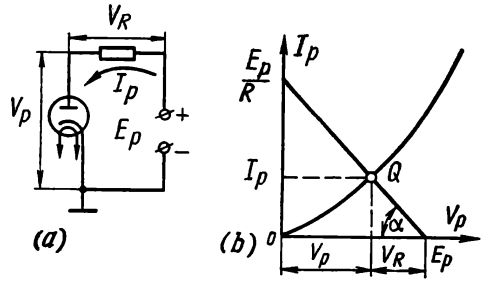


Fig. 2-10. (*a*) Operation of a vacuum diode with a load resistor in the plate circuit and (*b*) construction of the load line

the scale coefficients (according to the scales of voltage and current used), each having dimensions of its own.

Point Q, where the load line intersects the plate characteristic, represents the solution of Eq. (2-15). The projection of this point on the y-axis determines the current I_p flowing in the load circuit, and

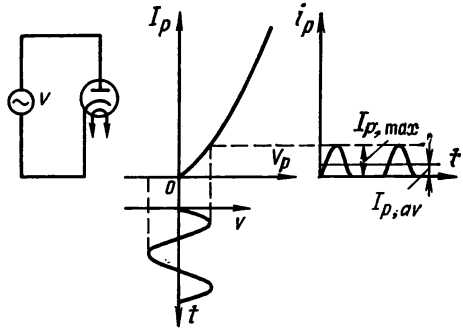


Fig. 2-11. A. C. equivalent circuit of a diode

its projection on the x-axis gives the plate voltage V_p and the load voltage V_L.

This graphical method is widely used in solving electronic circuits based on vacuum tubes.

One of the major properties of the diode is unidirectional conduction. Figure 2-11 illustrates operation of the diode when its electrodes are connected to a source of alternating voltage. In the circumstances, current flows through the diode only during the positive half-cycles of the sinusoidal voltage, that is, when the potential diffe-

rence between the plate and the cathode is positive. Therefore, current i_p is a sequence of pulses, which may be characterized by maximum instantaneous plate current, $I_{p,max}$, and average plate current, $I_{p,\,av}$, that is the constant component of the periodical pulse sequence (rectified current). During negative half-cycles, no current flows through the diode. Unidirectional conduction forms the basis for operation of the diode as a rectifier. Diode tubes specially designed for this purpose are called *rectifier* (or *rectifying*) *diodes.*

2-5. PARAMETERS AND RATINGS OF THE VACUUM DIODE

The performance of a diode can conveniently be described in terms of several variables, such as diode parameters, ratings, mechanical factors, etc. According to the designated function of a tube, the variables listed in a data sheet or a handbook may be modified. This section deals only with the basic variables characterizing the properties of a majority of diode tubes. An analysis of some variables requires detailed discussion of tube operation in special circuits. They are given in courses describing the operation of concrete electronic devices.

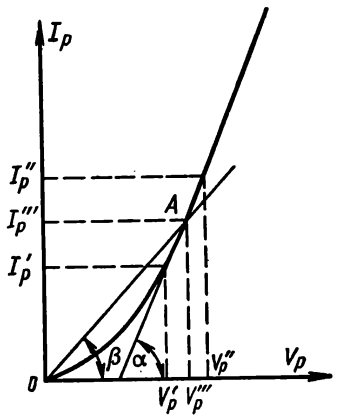


Fig. 2-12. Diode parameters

Diode Parameters. These characterize the performance of a tube under static conditions. The parameters related to the vacuum diode are plate conductance, g_p, dynamic plate resistance, R_p, and static or d.c. resistance, R_0.

The *plate conductance* of a vacuum diode is defined as the ratio of a slight change in plate current to the slight change in plate voltage that caused it. Since the plate characteristic of a diode is nonlinear, the plate conductance is not the same for different points of the characteristic.

For any point of the plate characteristic

$$g_p = dI_p/dV_p \tag{2-16}$$

As is seen, plate conductance is numerically equal to tan α, that is, the slope of the plate characteristic at a given point A (Fig. 2-12). The plate conductance of a vacuum diode can approximately be determined from its characteristic, as shown in Fig. 2-12, by finding the incremental change of plate current caused by an incremental change of plate voltage within the specified portion of the characteristic near point A, and by dividing the former by the latter:

$$g_p = (I_p'' - I_p')/(V_p'' - V_p') = \Delta I_p/\Delta V_p \tag{2-17}$$

Plate conductance has the same dimension as electric conductance (siemens), but since plate current is usually measured in milliamperes and plate voltage in volts, it is expressed in milliamperes per volt.

For an ideal diode, the plate conductance can be derived by differentiating Eq. (2-11) which expresses the three-halves power law

$$g_p = dI_p/dV_p = (3/2)GV_p^{1/2} = (\text{approx.})\ 3.5 \times 10^{-6}\,(Q_p/r_p^2)\,V_p^{1/2} \quad (2\text{-}18)$$

This expression is useful when designing diodes and establishing the relationship between the plate conductance and size of the tube electrodes. The plate conductance of various diodes, usually taken within the steepest region of the plate characteristic, ranges from one to several milliamperes per volt.

The *dynamic plate resistance of a vacuum diode*, symbolized R_p, is the reciprocal of the plate conductance g_p:

$$R_p = dV_p/dI_p = 1/g_p \quad (2\text{-}19)$$

At point A (Fig. 2-12) it may be approximately written as

$$R_p = (V_p'' - V_p')/(I_p'' - I_p') = \Delta V_p/\Delta I_p \quad (2\text{-}20)$$

The dynamic plate resistance which correlates changes in voltage and current may be regarded as the resistance which the diode offers to the alternating current flowing through it. For most diodes, the value of R_p ranges from tens of ohms to several kilohms.

When a direct current flows through a diode, of interest is not the ratio of an incremental change in current to the incremental change in voltage near some specified point, but the values of current in and voltage across the tube. In this case, the characteristic near this point is of no importance because the direct voltage and the respective current do not vary with time.

The *static* or *d.c. resistance of a vacuum diode* is the ratio of plate voltage to plate current. For example, the static resistance at point A in Fig. 2-12 is

$$R_0 = V_p'''/I_p''' \quad (2\text{-}21)$$

R_0 is numerically equal to the cotangent of the angle that the straight line drawn from the origin of coordinates to point A makes with the x-axis (cot β in Fig. 2-12). The value of R_0 may be either higher or lower than that of R_p, according to the region of the characteristic in which it is determined. In the rising part of the characteristic, $R_p < R_0$. For idealized characteristics obeying the three-halves power law, the relationship between R_p and R_0 may be derived with the help of Eqs. (2-12) and (2-18):

$$R_0 = V_p/I_p = V_p/GV_p^{3/2} = 1/GV_p^{1/2} = 3R_p/2 \quad (2\text{-}22)$$

Interelectrode Capacitance. The electrodes in a diode are insulated conductors separated by a small distance. Because of this, they

act as a small capacitor, called the *interelectrode capacitance.* Its value ranges from several picofarads in low-power diodes to tens of picofarads in high-power tubes. To write the interelectrode capacitance for a planar diode, we may use the expression for the capacitance of a plate-parallel capacitor

$$C = \varepsilon_0 Q/r_p \tag{2-23}$$

Here, r_p is the distance between the cathode and the plate and Q is the surface area of the electrode.

The above expression may be used to determine the interelectrode capacitance of a diode when its cathode is cold and there is no space charge. Such interelectrode capacitance is known as *cold capacitance.*

In real tubes, the capacitance formed by their electrodes is supplemented by the capacitance between the electrode leads. So, in fact, the cold interelectrode capacitance exceeds the value given by Eq. (2-23).

In an operating diode, the space charge is other than zero, and the potential distribution in the cathode-plate space is nonlinear.

As plate voltage V_p changes, the charges at the plate and cathode of a tube vary by different amounts due to the nonuniform distribution of the space charge in the interelectrode space. When a diode is operating in a circuit, it is important to know the plate current component due to the interelectrode capacitance. For this purpose, the interelectrode capacitance of an operating diode is defined as a certain differential (or dynamic) capacitance

$$C' = dq_p/dV_p \tag{2-24}$$

where

$$q_p = \varepsilon_0 Q \mathscr{E}_p \tag{2-25}$$

is the charge at the plate.

The field strength $\mathscr{E}_p$ near the plate surface may be found with the help of Eq. (2-8). On substituting $x = r_p$ and $V = V_p$ in this expression and dividing Eq. (2-9) by the new equation, we may write

$$V/V_p = (x/r_p)^{4/3} \tag{2-26}$$

After differentiating this expression with respect to x and substituting $x = r_p$ and $\mathscr{E} = \mathscr{E}_p$, we get

$$\mathscr{E}_p = (4/3)\ (V_p/r_p) \tag{2-27}$$

Substituting the above expression in Eq. (2-25) we may write according to Eq. (2-24)

$$C' = (4/3)\ (\varepsilon_0 Q/r_p) \tag{2-28}$$

This capacitance is known as *hot capacitance.* Due to the action of the space charge it is 4/3 times capacitance C.

Thus, so long as its cathode is cold, a vacuum diode has a cold capacitance C. When the cathode is not, the diode may be replaced by an a.c. equivalent circuit in which it is visualized as a resistor R_p

and a capacitor C connected in parallel. At low frequencies, the capacitive reactance, $X_C = 1/i\omega C'$, is many times the diode resistance, R_p, and its bypass effect is not felt. At high frequencies, the effect of interelectrode capacitances may be very significant (see Sec. 4-6).

In some circuits, the capacitance between the heater and the cathode, C_{hk}, is of considerable importance. The value of this capacitance for some types of diodes is stated in handbooks.

Bogey Values. As a rule, tube manufacturers specify what is known as *bogey values.* These are the average, or published, values for tube characteristics; these values apply to the operation of a tube under specified (reference) conditions. Reference conditions for diodes are filament voltage V_f and plate voltage V_p. Bogey values usually include nominal (rated) filament current I_f, maximum rectified plate current $I_{p,\,r}$, maximum instantaneous plate current $I_{p,\,max}$, initial plate current I_{p0} at $V_p = 0$, and some other quantities related to the purpose of a tube.

Ratings. These are values which establish either limiting capabilities or limiting conditions for a diode. As a rule, data sheets specify a *minimum* and a *maximum filament voltage.* As already noted, overheating of a cathode cannot be tolerated, because it would cut down the life of the cathode. On the other hand, if an oxide-coated cathode is operated at an insufficient filament voltage, the low temperature raises the volume resistance of the oxide layer, and the cathode may be overheated, too. Underheating is most dangerous in tubes with high plate currents. The areas on the cathode with an increased resistance are heated by the tube current more than other areas, so local overheating occurs.

Another rating is the *peak-inverse voltage* which sets the limit for the amplitude of the applied alternating voltage.

During the negative half-cycles of the applied alternating voltage, no current flows in a diode and all of the inverse voltage is applied between the plate and the cathode of the tube. If the peak-inverse voltage exceeds the safe value, the insulation between the plate and cathode leads may break down, or an arc-back may occur in the interelectrode space.

It is also important to know the *peak pulse plate current* and *maximum effective plate current* (Fig. 2-11). The limit for the pulse current may be set either by the emission current density (say, for tungsten cathodes) or by the operating conditions (for oxide-coated cathodes) under which overheating would cause no damage. The same factors in combination with the maximum safe value of plate dissipation limit the value of maximum effective plate current.

As already noted, the ratings include the *maximum plate dissipation.* When electrons contributing to plate current strike the plate, almost all of their kinetic energy is converted to the energy of ther-

mal oscillations within the crystal lattice of the plate material. This may considerably raise the plate temperature.

The power applied to the plate is

$$P_p = nmv^2/2 \qquad (2\text{-}29)$$

where n is the number of electrons striking the plate every second, m is the mass of an electron, and v is the velocity of an electron.

Using Eq. (2-2) we get

$$P_p = neV_p \qquad (2\text{-}30)$$

Here, the product ne is the electric charge delivered to the plate every second, that is the magnitude of plate current, I_p. Therefore, finally

$$P_p = I_p V_p \qquad (2\text{-}31)$$

For each type of vacuum diodes, data sheets specify the maximum plate dissipation, $P_{p,\max}$, which can easily be located on the plate characteristic (Fig. 2-13). The safe operation area of a tube is that bounded by this hyperbolic curve.

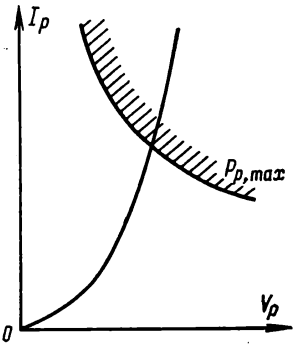


Fig. 2-13. Maximum plate dissipation

The plate gives up the absorbed energy mainly by radiation. According to the Stefan-Boltzman law, the total energy radiated per second is proportional to the absolute temperature to the fourth power. Thus, the value of P_p is limited by the maximum safe temperature which depends on a number of factors. In the first place, the plate temperature should not exceed the value at which the tube elements were degassed. Otherwise, the plate might give up gases and these would disturb the operation of the tube. Radiation from the hot plate can cause additional heating to the cathode. As already noted, cathode overheating increases emission from some spots, evaporates the active material from their surface and cuts down the life of the cathode.

The maximum plate dissipation at a specified temperature can be enhanced by increasing the surface area of the plate, by special treatment of its surface, and also by using forced-air or forced-water cooling.

2-6. CONSTRUCTION AND TYPES OF VACUUM DIODES

Receivers and low-power amplifiers use miniature and miniature diodes in glass envelopes (Fig. 2-14*a* and *b*). Their electrodes are of flat or cylindrical construction, the indirectly heated cathodes being of the oxide-coated type.

Rectifying vacuum diodes, handling larger amounts of power, use cylindrical electrodes. To avoid breakdown at high voltages, the plate lead is moved away from the cathode as much as possible and brought out through the top of the envelope (Fig. 2-14*c*).

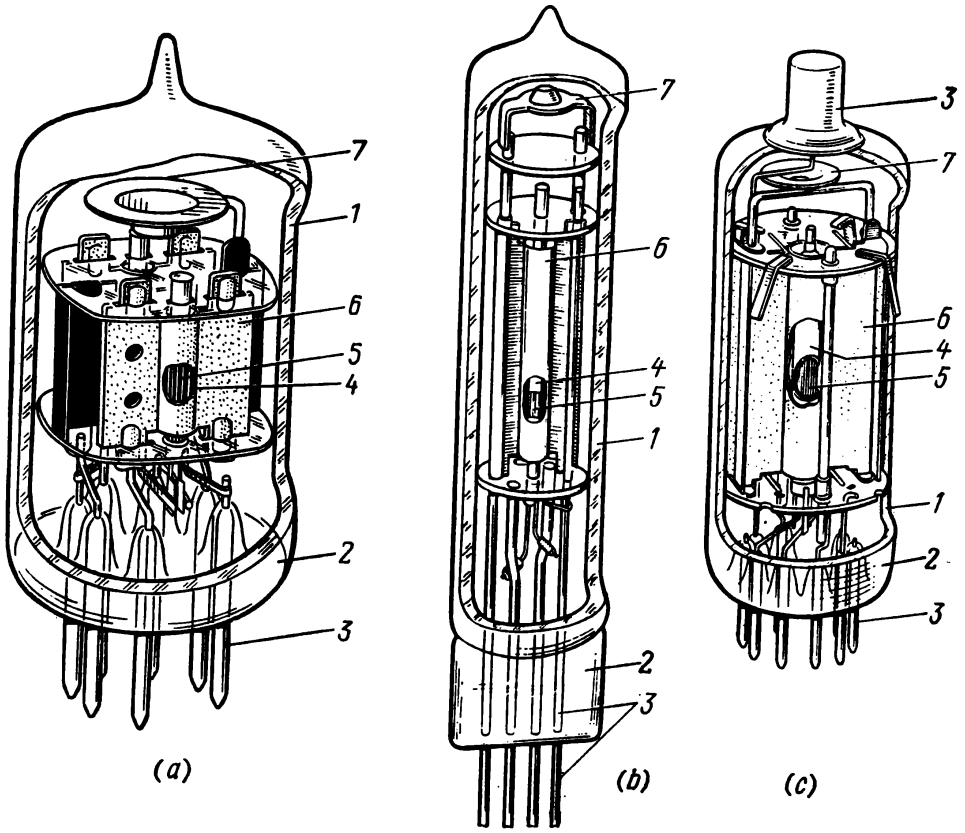


Fig. 2-14. Construction of some vacuum diodes
(*a*) miniature double diode; (*b*) microminiature diode; (c) rectifying diode; *1*—envelope; *2*—stem; *3*—pins and plate lead; *4*—cathode; *5*—heater; *6*—plate; *7*—getter

Some circuits use double diodes (two diodes in the same envelope). In most cases, vacuum diodes have recently been supplanted by crystal diodes.

Chapter Three

VACUUM TRIODES

3-1. PHYSICAL PROCESSES IN THE VACUUM TRIODE

Definition. Construction of the Triode. A vacuum triode is a three-electrode electron tube containing a plate, a cathode, and a control electrode (grid).

A sketch of a vacuum triode is shown in Fig. 3-1. In contrast to a diode where plate current is controlled by the electric field of the plate affecting the space charge near the cathode, in a triode, the flow of electrons is mainly controlled by the grid. The grid usually consists of a helix of fine wire coaxial with, and mounted close to, the cathode.

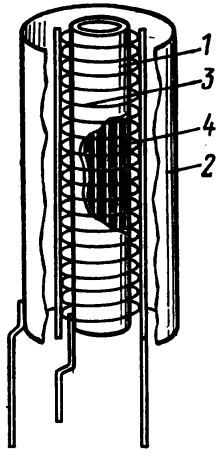


Fig. 3-1. Construction of a vacuum triode
1—cathode; *2*—plate; *3*—grid; *4*—heater

Figure 3-2 shows the connection of a triode in a circuit. The cathode potential is usually taken to be zero ($V_k = 0$), then the plate is at a positive potential, and the grid may be either at a positive or at a negative potential.

The circuit given in Fig. 3-2 is known as the grounded- (or common-) cathode connection. Other circuit configurations are the grounded- (common-) grid and the grounded- (common-) plate connection. In these connections, the grounded electrode is assumed to be at zero potential.

Potential Distribution in the Vacuum Triode. Figure 3-3 shows equipotential lines and potential distribution curves within a plane electrode triode at different grid voltages and a constant plate voltage. These curves are plotted for the cold tube, when the cathode is not heated and no electrons are emitted into the interelectrode space. Variations in the grid voltage markedly affect the electric field in the cathode-grid space; their effect on the field between the grid and the plate is weaker.

When the grid is sufficiently negative (Fig. 3-3*a*), a retarding ield in the cathode-grid space exists not only near the grid but also

between its wires. When the cathode is hot, (that is, emitting) electrons escaping from the cathode cannot overcome the retarding field and fall back on the cathode. As a result, the negative space charge grows in magnitude, and the potential in the cathode-grid space becomes still more negative. Electrons cannot reach the plate, and plate current is zero, $I_p = 0$ (the tube is said to be cut off). The lowest negative grid voltage (in absolute value) at which plate current becomes zero is called the *cut-off voltage*, $V_{g,\,c\text{-}o}$ (Fig. 3-3*b*).

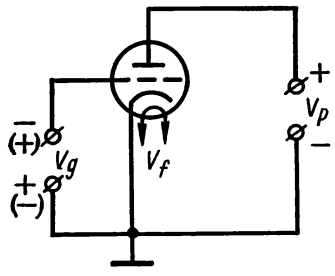


Fig. 3-2. Connection of a triode in a circuit

If the grid voltage is negative but more positive than the cut-off value, $|V_g| < |V_{g,\,c\text{-}o}|$ (Fig. 3-3*c*), the field between the grid wires becomes accelerating. When the cathode is emitting, electrons in this part of the space make for the plate and constitute plate current.

As the grid becomes less negative the retarding field is reduced, so that it remains only immediately at the grid wires (Fig. 3-3*c* and *d*), and a still larger number of electrons move towards the plate. When the grid is made positive (Fig. 3-3*e*), the accelerating field acts throughout the interelectrode space. The electrons moving near the grid wires are attracted to the grid and make up grid current I_g. Most electrons in the wire meshes move to the plate and constitute plate current, I_p. Thus, at $V_g > 0$, the stream of electrons travelling from the cathode and forming cathode current I_k is divided into two flows, the grid current and the plate current (Fig. 3-4)

$$I_k = I_g + I_p \tag{3-1}$$

But the surface area of the grid wires is much smaller than the surface area of the plate, so grid current is usually smaller than plate current.

From the foregoing it follows that variations in grid voltage considerably affect the field in the interelectrode space, especially that between the cathode and the grid. Thus, the potential barrier near the cathode and, as a consequence, the space charge are affected by a certain resultant field which is determined by plate and grid voltages.

Composite Controlling Voltage. To compare the effect of the plate and grid fields on the potential barrier near the cathode, it is customary to consider the field between the cathode and the grid as a field established by a certain equivalent voltage applied to the grid. In this method, one considers an imaginary diode consisting of the cathode of the original triode and a virtual plate having the same surface area and replacing the actual grid (Fig. 3-5). The fields

in the grid-cathode space of the original triode and the plate-cathode space of the equivalent diode are taken to be equivalent if the charges induced on the cathode surface of each tube are equal.

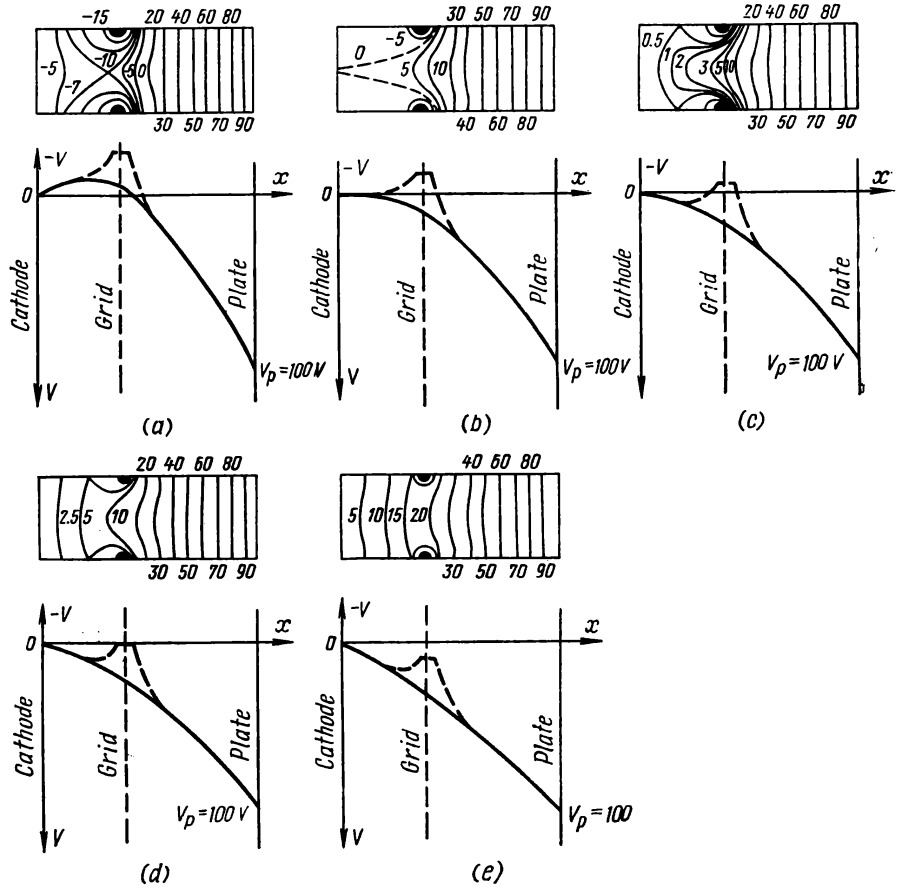


Fig. 3-3. Equipotential surfaces and potential distribution in a planar triode (solid lines — between grid wires; dashed lines — through a grid wire)
(*a*) $V_g = -25$ V; (*b*) $V_g = V_{g,c-o} = -12$ V; (*c*) $V_g = -6$ V; (*d*) $V_g = 0$; (*e*) $V_g = +10$ V

The charge induced on the diode cathode is

$$q_d = CV_d \tag{3-2}$$

Here, C is the plate-cathode capacitance of the diode and V_d is the composite controlling voltage, that is, the plate voltage of the equivalent diode.

The charge induced on the cathode of the triode depends on the plate and grid voltages and also on the plate-cathode and grid-cathode

capacitances (Fig. 3-5)

$$q_{tr} = C_{gk}V_g + C_{pk}V_p \tag{3-3}$$

where C_{gk} is the grid-cathode capacitance, C_{pk} is the plate-cathode capacitance, and V_g and V_p are the grid and plate voltages, respectively.

If the diode is equivalent to the triode in the sense defined above, the charges induced on the cathode surface of each tube are equal, that is, $q_d = q_{tr}$.

Hence,

$$CV_d = C_{gk}V_g + C_{pk}V_p \tag{3-4}$$

and

$$V_d = (C_{gk}/C)(V_g + C_{pk}V_p/C_{gk}) \tag{3-5}$$

We have replaced the grid by the virtual plate, so $C > C_{gk}$. Strictly speaking, $C \neq C_{gk} + C_{pk}$. However, to simplify the equations, we shall take it that C = (approx.) $C_{gk} + C_{pk}$.

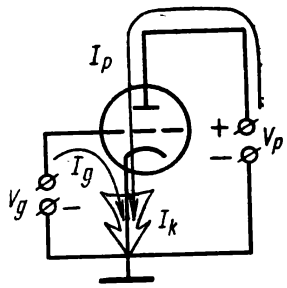


Fig. 3-4. Currents in a triode

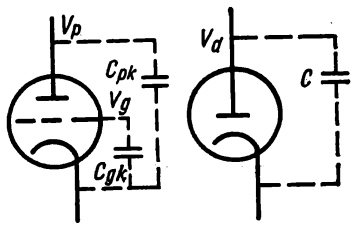


Fig. 3-5. Equivalent diode

Substituting this expression into Eq. (3-5) and putting

$$C_{pk}/C_{gk} = D \tag{3-6}$$

we get

$$V_d = (\text{approx.})\frac{1}{1+D}(V_g + DV_p) \tag{3-7}$$

The capacitance ratio, C_{pk}/C_{gk}, in a triode is generally much smaller than unity, so D may be neglected in comparison with the unity in denominator.

Then,

$$V_d = (\text{approx.})\ V_g + DV_p \tag{3-8}$$

Penetration of the Triode. As is seen from the above expression the composite controlling voltage directly depends on grid voltage V_g. As for plate voltage, it appears in Eq. (3-8) with a coefficient

$D < 1$ which shows that the plate voltage has a lower effect on the field in the cathode-grid space. The quantity D is called the *penetration factor* (or, simply, *penetration*) *of a tube* (also known as the *grip-through*)*. The penetration of a tube shows how much the plate field extends into the grid-cathode space, and takes account of the reduced effect of this field on the potential barrier near the cathode as compared with the grid field.

The effect of plate voltage is much lower because (1) the plate is farther away from the cathode than the grid, and (2) which is more important, the grid screens the cathode from the plate, and the lines of force from the plate penetrate into the grid-cathode space but little. The screening effect of the plate field considerably depends on the geometry of the electrodes (mainly the grid), which is confirmed by Eq. (3-6) defining D as the ratio of capacitances, C_{pk}/C_{gk}. As the spacing between the grid wires decreases, the number of lines of force penetrating through the grid also decreases and so does D because a decrease in grid-wire spacing brings about a rise in C_{gk} and a fall in C_{pk}.

The Three-Halves Power Law for the Vacuum Triode. Using the concept of the composite controlling voltage, we may apply the three-halves power law to the vacuum triode. This can be demonstrated as follows. For the equivalent diode, this law may be written similarly to Eq. (2-12)

$$I_{p,e.d} = G'V_d^{3/2} \tag{3-9}$$

Because of the equivalence of the diode and the triode, the electron flows from their cathodes should be equal. In the equivalent diode, this flow corresponds to plate current, $I_{p,\,e.\,d}$, and in the triode, to cathode current, I_k.

Recalling Eqs. (3-1) and (3-7), we may write Eq. (3-9) as

$$I_k = I_p + I_g = G'\left[\frac{1}{1+D}(V_g + DV_p)\right]^{3/2} \tag{3-10}$$

where, for a cylindrical-electrode triode,

$$G' = 2.33 \times 10^{-6} Q_p / r_p r_g \beta^2 \tag{3-11}$$

Here, r_g is the radius of the grid, and r_p is the radius of the plate.

For planar tubes, $\beta^2 = 1$, and the product $r_p r_g$ in the denominator should be replaced by r_{kg}^2 which is the square of the distance from the cathode to the grid.

It should be noted that, as with the diode, the relationship between the cathode current and the voltages at tube electrodes given by Eq. (3-10), differs from real static characteristics (this will be

* From the German "Durchgriff".— *Translator's note.*

discussed in more detail in Sec. 3-2). Therefore, the three-halves power law is less applicable to the vacuum triode than to the vacuum diode. In practice, use is made of other methods of approximation to real triode characteristics (see Chapter 5).

Cut-off Voltage. Using Eq. (3-10), we can derive a formula for the cut-off voltage of a triode, $V_{g,\,c\text{-}o}$, that is, the voltage at which plate current becomes zero. If we neglect D in Eq. (3-10), because $D \ll 1$, we get

$$I_k = G' (V_g + DV_p)^{3/2} = G' V_d^{3/2} \tag{3-12}$$

On putting $I_g = 0$, we may write

$$I_p = G' (V_g + DV_p)^{3/2} \tag{3-13}$$

When a voltage equal to the cut-off value is applied to the grid, plate current is zero. So

$$0 = G' (V_{g,\,c\text{-}o} + DV_p)^{3/2} \tag{3-14}$$

Since G' is non-zero, then

$$V_{g,\,c\text{-}o} + DV_p = 0 \tag{3-15}$$

Hence

$$-V_{g,\,c\text{-}o} = DV_p \tag{3-16}$$

Thus, the cut-off voltage of a tube increases (in absolute value) with increasing penetration and plate voltage of the tube. However, it should be kept in mind that the cut-off voltage determined by Eq. (3-16) is somewhat smaller than the real value, because it is based on the three-halves power law which does not take into account some physical processes in a tube.

The Physical Meaning of the Penetration Factor. Using the three-halves power law, we can derive a definition of the tube penetration differing from Eq. (3-6).

For this purpose, let us find the differential dI_k with the help of Eq. (3-12)

$$dI_k = (3/2)\, G' V_d^{1/2} (dV_g + D\, dV_p) \tag{3-17}$$

where dV_g and dV_p are the independent increments in the grid and plate voltage, respectively.

This suggests that cathode current can be controlled by varying V_g and V_p. If we reduce the negative grid voltage by dV_g, we shall produce some incremental change in the cathode current, dI_k. This increment may be then compensated for by reducing the plate voltage by dV_p. So, in the end, the cathode current will attain its former value, and dI_k will be zero. Then, according to Eq. (3-17),

$$dV_g + D\, dV_p = 0$$

whence

$$D = -\, dV_g/dV_p \text{ at } I_k = \text{const} \tag{3-18}$$

This expression shows more clearly the physical meaning of the tube penetration factor as a quantity characterizing the relative effect which plate and grid voltages have on the potential barrier near the cathode and, therefore, the cathode current. The "minus" sign in Eq. (3-18) shows that in order to maintain a constant cathode current, increments in plate and grid voltages should differ in sign, that is, a fall in the negative grid voltage should be accompanied by a fall in the positive plate voltage. Then the penetration factor will always be a positive quantity.

Current Partition. As already stated, at $V_g > 0$, the flow of electrons which have overcome the potential barrier is divided into

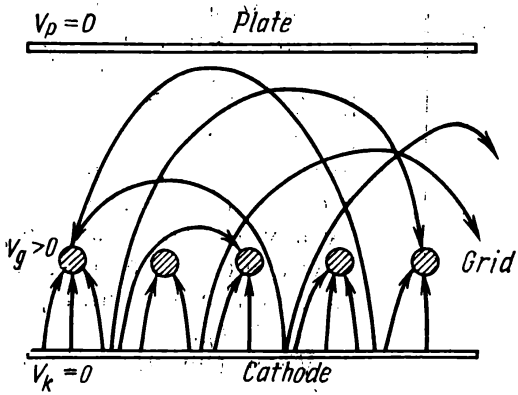


Fig. 3-6. Electrons flying back to the grid

two streams: some electrons move to the plate (the plate current) and the others move to the grid (the grid current). This division of the cathode current carried by the total flow of electrons between the various electrodes or circuits of the tube is called *current partition.*

Quantitatively, this is described in terms of the current partition ratio

$$k_{cp} = I_p / I_g \tag{3-19}$$

According to Eq. (3-1), $I_k = I_p + I_g$, so

$$I_p = k_{cp} I_k / (k_{cp} + 1) \tag{3-20}$$

$$I_g = I_k / (k_{cp} + 1) \tag{3-21}$$

The values of I_p and I_g and, therefore, the current partition ratio k_{cp}, depend on V_p and V_g.

Now let us discuss the motion of electrons in a triode at different relative magnitudes of plate and grid voltages.

Electron Flyback. Let the grid of a triode be at a positive potential, and plate voltage vary from zero to a value which exceeds the grid voltage. At $V_g > 0$ and $V_p = 0$, the electrons moving through the grid wires are affected by the positive field of the grid, and their paths are bent (Fig. 3-6). The grid has its maximum effect on the

electrons travelling near its wires. Some electrons move to the grid directly; others, on passing through the grid, are slowed down by the retarding field of the grid, brought to rest and pulled back onto the grid in curved paths. The plate current is zero, and all electrons that have overcome the potential barrier near the cathode fall on the grid.

The situation is different when a small positive voltage, $V_p < V_g$, is applied to the plate. A small proportion of electrons moving through the grid wires proceed to the plate. The energy of these electrons is sufficient to overcome the plate field whose retarding effect depends on the potential difference $V_p - V_g$. A considerable number of electrons travelling near the grid, accelerated by its field towards the grid wires, travel along bent path and fall back on the grid. In comparison with the previous case, the number of these electrons is reduced. As a consequence, the grid current decreases by a value equal to the plate current, so now the cathode current is equal to the sum of the grid and plate currents, $I_k = I_g + I_p$.

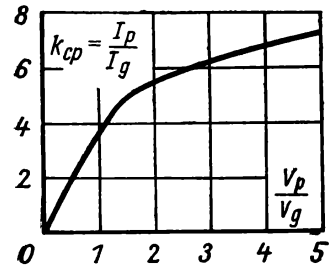


Fig. 3-7. Current partition ratio as a function of V_p/V_g

When the plate is made still more positive, the plate current swiftly rises, and the grid current falls in proportion. The electrons moving between the grid and the plate are directly affected by the plate field, because the grid has no screening effect in this part of the interelectrode space. The current partition ratio, $k_{cp} = I_p/I_g$, quickly rises (Fig. 3-7).

The relative values of V_p and V_g at which electrons are pulled from the grid-plate space back to the grid correspond to a condition known as *electron flyback*. Referring to Fig. 3-7, electron flyback usually takes place in a tube at $V_p/V_g \leqslant 1$, that is, when a retarding field exists in the grid-plate space.

Electron Interception. As the plate voltage increases further, no more flyback occurs. Now the positive field of the grid only bends the paths of the electrons moving to the plate, and the grid current is constituted only by the electrons directly intercepted by the grid wires. This condition is known as *electron interception*.

Under conditions of electron interception, the relative values of plate and grid currents are mainly determined by the ratio of the surface areas of these electrodes. Since the surface area of the grid wires is generally much smaller than the effective area of the plate, plate current is considerably higher than grid current.

With electron interception, the current partition ratio is mainly controlled by the plate current, I_p, because the grid current decreases insignificantly as V_p increases. The plate current rises owing to

the effect of the plate voltage on the field between the cathode and the grid. Since this effect is weakened by the screening action of the grid, the current partition ratio under conditions of electron interception grows more slowly with increasing V_p/V_g (Fig. 3-7).

The transition from flyback to interception also depends on the construction of the grid and the electron flow density.

Under conditions of electron flyback, the current partition ratio mainly depends on the relative magnitude of the cathode-grid and cathode-plate distances, and the grid wire spacing, t. In electron interception, k_{cp} is chiefly determined by the grid space factor, δ/t (where δ is the wire diameter). As already noted, the transition from electron flyback to electron interception may occur at different values of V_p/V_g, according to the tube electrode construction. For most triodes, this value ranges from 0.2 to 1.2.

Current partition in a tube may also markedly be affected by secondary emission from the plate or the grid. Sometimes, this effect is aggravated if the plate or the grid is contaminated by particles of the cathode activator.

Secondary electrons emitted by the plate or the grid may be accelerated by strong fields towards the other electrodes, for example, from the plate to the grid, if the grid is more positive than the plate. The flow of secondary electrons is opposite to that of primary electrons, so plate current is reduced.

3-2. STATIC CHARACTERISTICS OF THE VACUUM TRIODE

In the previous section, we learned that the electron flow in a vacuum triode is controlled by electric fields which are established by the external supply sources connected to the tube electrodes.

Generally, the total cathode current in a triode may divide into a plate and a grid current. Because of this, it is necessary to consider the relationships between each of the currents and tube electrode voltages separately. In electronic circuits, a control signal is usually applied to the grid, so the grid circuit is input; an output signal is taken from the plate, so the plate circuit is output. Hence, the most important functions are

$$I_p = f(V_g;\ V_p) \tag{3-22}$$

$$I_g = \varphi(V_g;\ V_p) \tag{3-23}$$

The effect of the electric fields in a triode on the electron flow and, as a consequence, on the electrode currents, is rather complex in character. It depends on the electrode construction, relative magnitude of voltage, and some other factors. Even an approximate analytic relationship between cathode current and electrode volta-

ges (the three-halves power law) shows that the triode does not obey Ohm's law.

The relationships expressed by Eqs. (3-22) and (3-23) are nonlinear. Each relation which is a function of two independent variables may be written in the total differential form

$$dI_p = (\partial I_p/\partial V_g)dV_g + (\partial I_p/\partial V_p)dV_p \quad (3\text{-}24)$$

$$dI_g = (\partial I_g/\partial V_g)dV_g + (\partial I_g/\partial V_p)dV_p \quad (3\text{-}25)$$

The increments dV_g and dV_p may be regarded as very small changes in direct voltages when the tube operates at no load. In operation under load, they may be replaced by alternating signal voltages if they are sufficiently small for the relationship to be deemed linear within these limits.

Let us discuss the operation of a triode at no load. On putting one of the voltages constant, we may represent the functions (3-22) and (3-23) by four relationships known as the *static characteristics of the triode*, namely

— the plate-grid (transfer or mutual) characteristic

$$I_p = f_1(V_g) \qquad V_p = \text{const} \quad (3\text{-}26)$$

— the plate characteristic

$$I_p = f_2(V_p) \qquad V_g = \text{const} \quad (3\text{-}27)$$

— the grid characteristic

$$I_g = \varphi_1(V_g) \qquad V_p = \text{const} \quad (3\text{-}28)$$

— the grid-plate characteristic

$$I_g = \varphi_2(V_p) \qquad V_g = \text{const} \quad (3\text{-}29)$$

These characteristics tell how the plate or the grid current is related to the voltage at one of these electrodes, if the voltage at the other is kept constant.

The four families of characteristics may differ from triode to triode according to tube construction, purpose, and so on, but the general appearance is the same for any vacuum triode.

Plate-Grid Characteristics. Figure 3-8 shows the plate-grid characteristic curves of a vacuum triode. Even in the presence of a high positive voltage at the plate, plate current begins to flow only when the grid voltage is $|-V_g| < |-V_{g,c\text{-}o}|$. When the grid is negative but more positive than $V_{g,c\text{-}o}$, the plate current in the tube is zero and the tube is cut off.

The value of cut-off voltage given by Eq. (3-16) depends on both plate voltage and the construction features of a triode that affect the tube penetration factor, D. For example, in tubes where the grid wires are spaced close apart, and the capacitance C_{pk} and the penetration factor D are small, the cut-off voltage is low and the

tail of their characteristics is very short. At the same plate voltage, the tube will be cut off by a less negative $V_{g,c\text{-}o}$.

A comparison of two characteristics in this familly would show that increasing plate voltage makes the characteristic move to the left (it is said to have a long tail), because according to Eq. (3-16), the cut-off voltage rises, too.

Figure 3-8 also shows that the relationship $I_p = f_1(V_g)$ is nonlinear. As the grid becomes less negative (rises above the cut-off level), the plate current increases, at first slowly (the lower-curved portion),

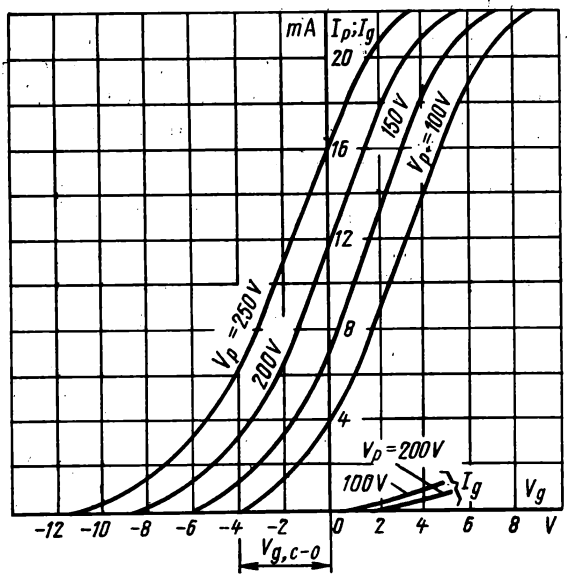


Fig. 3-8. Families of plate-grid and grid characteristics of a triode

then faster, and the characteristic becomes almost linear. In the region of positive values of V_g, the grid draws current and the curve is more gradual which is an indication that the growth of plate current is slowed down.

It is obvious that this character of the $I_p = f_1(V_g)$ function cannot satisfactorily be described by the three-halves power law. The point is that Eq. (3-10) is written for cathode current, $I_k = I_p + I_g$, whereas the plate-grid characteristics represent variations in plate current. So, the three-halves power law applies only to that part of the plate-grid characteristic where $I_g = 0$, that is, to the region of negative values of V_g. But even at a negative V_g, the actual plate-grid characteristic markedly differs from the three-halves power law. For example, the lower curved portion of the actual characteristic is more gradual than it ought to be according to Eq. (3-10).

This and other differences are explained not only by the causes described in Sec. 2-3 for the vacuum cathode and applicable to a great extent to the triode, but also by some other factors peculiar to grid-controlled tubes.

The most important factor is the nonuniformity of the resultant field near the cathode. This can easily be seen in Fig. 3-3. At a negative grid voltage close to the cut-off value, the retarding field is established only near those parts of the cathode which are located directly under the grid wires (Fig. 3-3*c*). Due to the effect of the plate, the resultant field between the grid wires can be accelerating. As a result, electrons will leave these parts of the cathode, even though the grid is very negative. The bottom portion of the actual characteristic becomes more gradual, and the cut-off voltage of the triode more negative. This effect, known as the *island effect*, becomes more pronounced when the grid wires are spaced wider apart and the grid is located closer to the cathode.

At $V_g > 0$, the difference between the actual plate-grid characteristic and the three-halves power law is still more striking. Plate current becomes smaller than cathode current because the electron flow is divided into plate and grid current. As V_g grows more positive, grid current rises, and the growth of plate current is slowed down. However, at a low positive grid voltage such that $V_g \ll V_p$, the tube operates under conditions of electron interception, I_g is low, and I_p insignificantly differs from I_k. As V_g increases, this difference becomes more pronounced.

Grid Characteristics. Graphically, the $I_g = \varphi_1(V_g)$ functions are represented by a series of divergent curves (Fig. 3-8). Here, at the same V_g, grid current decreases with increasing plate current, because the current partition ratio, $k_{cp} = I_p/I_g$, rises quicker than the cathode current, I_k.

If, in measuring grid characteristics, we connect a sufficiently sensitive instrument in the grid circuit, we shall see that the $I_g = \varphi_1(V_g)$ curves do not always start from the origin of coordinates. In some cases, grid current can be registered already at low negative grid voltages (Fig. 3-9*a*), in other cases, it appears only at positive grid voltages of 0.2 to 0.5 V. In the latter case, the grid characteristics move to the right of the origin of coordinates (Fig. 3-9*b*).

In the former case, the position of the $I_g = \varphi_1(V_g)$ curve may be characterized by the initial grid current I_{g0} at $V_g = 0$ and the grid cut-off voltage $V_{g.c-o}$. In some triode applications, grid current at negative grid voltages is undesirable because it would result in a reduced input resistance of the tube. I_{g0} is usually low, being tens of microamperes in low-power tubes and from 100 to 300 μA in high-power tubes.

There are several factors which cause the grid characteristic to shift from the origin of coordinates. To begin with, we shall discuss

those for actual characteristics of the vacuum diode, considered in Sec. 2-3. At $V_g < 0$, the grid draws some current owing to the fastest electrons emitted by the cathode. Their initial energy may be sufficient to overcome the low retarding field near the grid, and the electrons are intercepted by the grid wires.

Most often, grid characteristics are shifted to the right due to the contact-potential difference between the grid and the cathode. As a rule, the grid is made of molybdenum whose work function is

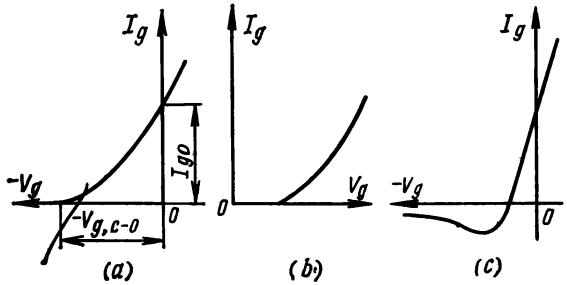


Fig. 3-9. "Left-hand" and "right-hand" characteristics

considerably higher than that of the cathode coating. This brings about a contact-potential difference equal to the difference between the work functions. So, even at $V_g = 0$, a low retarding field exists in the cathode-grid space.

The grid circuit may also draw a current, although a small one, called, *gas current*, which usually appears in tubes with imperfect vacuum. The point is that electrons accelerated by the tube may be sufficient to ionize molecules of the residual gas in the envelope. Then the electrons produced by the ionization add to the plate current, the positive ions are attracted by the negative grid and draw electrons from the external circuit. This gas current usually rises with falling negative grid current, because, as this takes place, the flow of primary electrons increases. The value of gas current gives a measure of the pressure of the residual gas.

The gas current may also be caused by electron emission from the grid wires. It depends on how much the grid is contaminated by the vaporized cathode material and heated by the cathode. To eliminate these factors, heat abstraction from the grid is improved by using massive supports, and the grid wires are gold-plated to increase the diffusion of deposited atoms into the coating and reduce emission from the grid wires.

The above phenomena account for the shape of the grid current characteristic shown in Fig. 3-9*c*.

Plate Characteristics. Graphically, the $I_p = f_2(V_p)$ function is represented by a family of monotonically increasing curves

(Fig. 3-10). When the grid is negative ($V_g < 0$), grid current is zero and plate current is equal to cathode current. Plate current begins to flow at a certain positive plate voltage, V_{p0}, which can be determined by Eq. (3-13), assuming that $I_p = 0$

$$V_{p0} = -V_g/D \qquad (3\text{-}30)$$

When the grid becomes more negative, V_{p0} increases, that is, the characteristic moves to the right. At $V_g > 0$, the plate characteristics start from the origin of coordinates and have a typical knee. The initial steeply rising portion of each curve corresponds to elect-

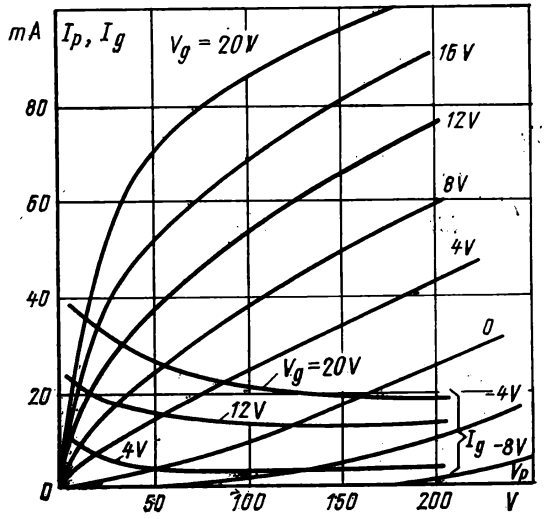


Fig. 3-10. Families of plate and grid-plate characteristics of a triode

ron flyback. Under conditions of electron flyback, the electrons returning to the grid form a space charge in the grid-plate space. The plate field, not screened by the grid, directly affects the motion of these electrons. As a result, plate current sharply rises at an insignificant increase in V_p. When plate voltage rises further, the grid begins to intercept electrons. Now I_p increases with increasing plate voltage because the plate voltage affects the space charge near the cathode. This effect of the plate voltage is weakened by the screening effect of the grid, and, naturally, the plate current grows more slowly; the curves now have a slope almost the same as in the case of $V_g < 0$, when the tube operates without grid current.

Grid-Plate Characteristics. In the $I_g = \varphi_2(V_p)$ characteristic curves (Fig. 3-10), the difference in the effect that plate voltage has on grid current under conditions of electron flyback and interception is also seen distinctly. When plate voltage varies from zero to several tens of volts, grid current decreases rather quickly from the value

equal to cathode current (electron flyback); as V_p rises still further, grid current falls slowly (electron interception). As the grid becomes less positive, the $I_g = \varphi_2 (V_p)$ curve shifts downwards because the composite voltage and cathode current decrease.

All characteristics discussed in this section are interrelated. For example, a family of plate characteristics may easily be plotted from

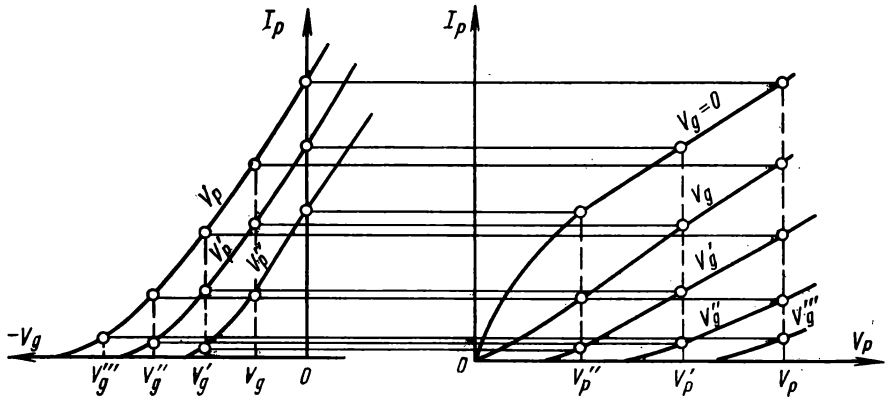


Fig. 3-11. Plotting a family of plate characteristics from a family of plate-grid characteristics

a family of plate-grid characteristics by transferring the points of the latter into the I_p/V_p coordinates as shown in Fig. 3-11.

3-3. PARAMETERS AND RATINGS OF THE VACUUM TRIODE

Static Parameters. These are constants describing the operation of a tube under static (no-load) conditions. They tell how effective the voltages at the triode electrodes are in controlling the currents in the respective leads. Let us first discuss the parameters of the plate circuit using Eq. (3-24).

$$dI_p = (\partial I_p/\partial V_g)dV_g + (\partial I_p/\partial V_p)dV_p$$

If we replace the partial derivatives $(\partial I_p/\partial V_g) = y_{pg}$ and $(\partial I_p/\partial V_p) = y_{pp}$, we may rewrite the above equation as

$$dI_p = y_{pg}\,dV_g + y_{pp}\,dV_p \qquad (3\text{-}31)$$

In the theory of two-ports (or four-terminal networks), the coefficients y_{pg} and y_{pp}, which have the dimension of admittance, are called *characteristic admittances* or *y-parameters*. Each characterizes the relationship between the voltage at one port and the current at the other. As applied to tubes, they describe the effect of the voltage at one electrode on plate current, with the voltages at the other

electrodes held constant. Accordingly, for tubes they are designated and named differently.

Transconductance. Assuming that V_p = constant and $dV_p = 0$, we get from Eq. (3-31)

$$y_{pg} = dI_p / dV_g \tag{3-32}$$

This quantity is called the *transconductance* (or *mutual conductance*) *of a triode* and symbolized g_m

$$g_m = \frac{dI_p}{dV_g}\bigg|_{V_p=\text{const}} \tag{3-33}$$

Transconductance is seen to be the rate of change of plate current with grid voltage, if plate voltage remains constant. Although the unit of transconductance is the siemens, more often it is expressed in mA/V.

Differentiating Eq. (3-13) and taking into consideration Eq. (3-8) we obtain an analytical expression for the transconductance of a triode with cylindrical electrodes

$$g_m = \partial I_p / \partial V_g = (3/2)\, G' V_d^{1/2}$$
$$= (\text{approx.})\, 3.5 \times 10^{-6} (Q_p / r_p r_g \beta^2) (V_g + DV_p)^{1/2} \tag{3-34}$$

For a planar triode, $\beta^2 = 1$ and $r_p r_g = r_{kg}^2$.

Equation (3-34) suggests that the transconductance increases with increasing effective area of the plate and decreasing distances between the electrodes.

The transconductance can most effectively be raised by reducing r_g or r_{kg}. In triodes with a high transconductance, this distance ranges from 60 to 100 μm. At a small r_g, the grid wires are spaced closer apart so as to eliminate the "islands" near the cathode (see Sec. 3-2) which make the grid less effective in its control of plate current and, as a consequence, bring down the transconductance g_m.

Data sheets usually specify transconductance for nominal (rated) values of V_f, V_g and V_p. According to the purpose of a tube, the transconductance varies from 1 to 50 mA/V. Most low-power triodes have the transconductance ranging from 1 to 10 mA/V.

A.C. (Dynamic) Plate Resistance. Assuming V_g = const and, therefore $dV_g = 0$, we obtain from Eq. (3-31)

$$y_{pp} = dI_p / dV_p \tag{3-35}$$

The reciprocal of y_{pp} is called the *dynamic plate resistance* of a triode (also known as the *a.c. plate resistance*)

$$R_p = \frac{dV_p}{dI_p}\bigg|_{V_g=\text{const}} \tag{3-36}$$

The dynamic plate resistance tells how plate voltage varies with plate current, if grid voltage is kept constant. As already noted, R_p

is the resistance that the tube presents to an alternating current. It gives a measure of variations in plate current caused by plate voltage with allowance for $I_p = f_2(V_p)$ at a certain operating point.

The value of the dynamic plate resistance depends on the mode of tube operation and electrode construction. In tubes with grid wires spaced close apart, R_p is high because plate voltage has a smaller effect on the space-charge region near the cathode. R_p also rises with increasing cathode-plate distance because this weakens the effect of plate voltage, too.

According to the type of triode and the mode of operation, the dynamic plate resistance ranges from 1 to 100 kilohms.

D.C. Resistance. Along with the dynamic plate resistance of a tube, which is one of the main tube parameters, it is sometimes convenient to speak of its performance in terms of the *d.c.* (or *static*) *resistance*, R_0. The difference between the static and dynamic plate resistance was discussed in Sec. 2-5: R_0 is the ratio of full values of plate voltage and current, but not their increments

$$R_0 = V_p/I_p \tag{3-37}$$

As a rule, the values of R_p and R_0 differ for different modes of operation.

No-Load Amplification Factor. The no-load parameters considered above — transconductance, g_m, and dynamic plate resistance, R_p — determine separately the effect of grid voltage and plate voltage on plate current, respectively. As follows from Eq. (3-13), variations in V_g affect I_p more than those in V_p.

The parameter which shows how effective these voltage are in controlling plate current in comparison with each other is called the no-load amplification factor (or, simply, the amplification factor), designated μ

$$\mu = -\left.\frac{dV_p}{dV_g}\right|_{I_p=\text{const}} \tag{3-38}$$

Why we require the plate current to be constant can be explained as follows. A change of dV_g in grid voltage brings about a change in plate current. To determine μ, plate voltage should be changed by dV_p so that plate current will take on the previous value. The "minus" sign shows that the effect of a change in plate voltage on plate current should be compensated for by an opposite change in grid voltage.

The amplification factor is a dimensionless quantity. For present-day triodes, it ranges from several unities to 100.

It should be emphasized that the amplification factor as well as the other two parameters, g_m and R_p, shows the effect of tube voltages on plate rather than cathode current. Accordingly, if in a triode $I_g > 0$ and $I_p \neq I_k$, only increments in plate current must be measu-

red when finding g_m and R_p, and plate current must be maintained constant when finding μ.

In contrast to μ, the penetration factor of a tube shows how effective V_g and V_p are in controlling cathode current (see Eq. (3-18))

$$D = -\frac{dV_g}{dV_p}\bigg|_{I_k=\text{const}}$$

The comparison of Eqs. (3-18) and (3-38) shows that, if $I_g = 0$, the amplification factor is the reciprocal of the tube penetration. To demonstrate, if $I_g = 0$, we can write $I_p = \text{const}$ instead of $I_k = \text{const}$ in Eq. (3-18), then

$$\mu = 1/D \tag{3-39}$$

However, it should be kept in mind that the above expression holds only if the grid draws no current ($I_p = I_k$).

Relation between Tube Parameters. The no-load parameters, g_m and R_p, have been derived on the basis of Eq. (3-31). Substituting the expressions for g_m and R_p into this equation, we get

$$dI_p = g_m dV_g + (1/R_p)\, dV_p \tag{3-40}$$

If $I_p = \text{const}$ or $dI_p = 0$, then

$$g_m R_p = \mu \tag{3-41}$$

This equation, sometimes called the *tube equation*, relates the three tube constants and is often used in circuit calculations.

Taking into consideration Eq. (3-39), we may re-write Eq. (3-41) as follows

$$g_m R_p D = 1 \tag{3-42}$$

This equation holds only in the absence of grid current as does Eq. (3-39).

No-Load Parameters of the Grid Circuit. Along with the plate circuit parameters, use is sometimes made of the grid circuit no-load parameters. These are determined by the coefficients of the total differential for grid current, Eq. (3-25)

$$dI_g = (\partial I_g/\partial V_g)\, dV_g + (\partial I_g/\partial V_p) dV_p$$

On replacing the partial derivatives by suitable y-parameters, we may re-write this expression as

$$dI_g = y_{gg}\, dV_g + y_{gp}\, dV_p \tag{3-43}$$

On putting $V_p = \text{const}$ or, which is the same, $dV_p = 0$, we get

$$y_{gg} = dI_g/dV_g \tag{3-44}$$

The quantity y_{gg} is called the *input conductance* of a tube, because, in most cases of triode operation, the grid-cathode circuit serves as the input to which the incoming signal is applied. Its reciprocal,

$R_{ig} = 1/y_{gg}$, is known as the *input resistance* or the *grid dynamic resistance of a tube.*

At $V_g < 0$, there is no grid current flowing, so theoretically the input resistance is infinite. In practice, however, it is finite, though very high ($R_{ig} \geqslant 10^7$ to 10^8 ohms), because current may leak through the insulation between the grid and cathode circuits.

If $V_g > 0$ and the grid draws current ($I_g \neq 0$), the input conductance y_{gg} is seen to be the rate of change of grid current with grid voltage, if plate voltage is held constant

$$g_g = \frac{dI_g}{dV_g}\bigg|_{V_p=\text{const}} \tag{3-45}$$

The special name for this quantity is the *grid conductance of a triode.*

If we take grid voltage to be constant (V_g = const or $dV_g = 0$), we may write on the basis of Eq. (3-43)

$$y_{gp} = dI_g/dV_p \tag{3-46}$$

This quantity characterizing the effect of plate (output) voltage in controlling grid (input) current is termed the *reverse transconductance.* At no-load, when $V_g < 0$ and $I_g = 0$, $y_{gp} = 0$.

At $V_g > 0$ ($I_g \neq 0$), the reverse transconductance is seen to be the rate of change of grid current with plate voltage, if grid voltage is held constant

$$g_{gp} = \frac{dI_g}{dV_p}\bigg|_{V_g=\text{const}} \tag{3-47}$$

The parameters g_g, g_{gp}, and g_m, are usually expressed in milliamperes per volt. At $V_g < 0$, I_g is zero and, as a consequence, $g_g = 0$ and $g_{gp} = 0$. For $V_g > 0$, g_g is positive and g_{gp} is negative, because grid current decreases with increasing V_p (see Fig. 3-10).

Determining No-Load Parameters from Characteristics. Tube manuals and data sheets specify the values of no-load parameters obtained under nominal operating conditions, that is, at certain values of V_g and V_p. These two quantities uniquely locate the operating (or quiescent) point on any family of characteristic curves. However, as is seen in Figs. 3-8 and 3-10, the static characteristics of a triode are nonlinear, so, for any other operating point (that is for V_g and V_p other than nominal) the values of the no-load parameters will be different. Because of this, in circuit calculations, the values of no-load parameters are determined from families of static characteristics.

Each of the tube parameters may be shown graphically on a family of static characteristic curves. Let grid and plate voltages have the values V_g' and V_p', respectively. Together they locate a certain quies-

cent point Q; at $V_g > 0$, this point may be plotted on any family of characteristic curves (Figs. 3-12 and 3-13).

On drawing the tangents to the characteristics at these points, we may define the no-load parameters of the grid and plate circuits in terms of the following trigonometrical functions

$$g_m = (m/n) \tan \alpha \tag{3-48}$$

$$R_p = (p/q) \cot \beta \tag{3-49}$$

$$g_g = (m/n) \tan \gamma \tag{3-50}$$

$$g_{gp} = (q/p) \tan \delta \tag{3-51}$$

Here, m, n, p and q are the scaling factors.

However, it is usually very difficult to measure the angles accurately. So, in practice, use is made of less accurate but more conven-

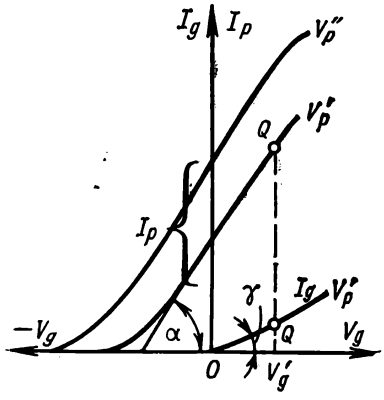


Fig. 3-12. Determining g_m and g_g from characteristics

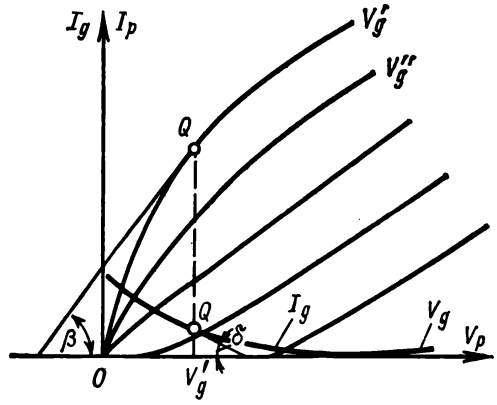


Fig. 3-13. Determining R_p and g_{gp} from characteristics

ient characteristic triangles. We choose the Q-point on a family of plate-grid characteristics (Fig. 3-14) and draw through it two lines parallel to the axes of coordinates and corresponding to, say, $I_p =$ $=$ const and $V_g =$ const. That $V_p =$ constant follows from the definition of grid-plate characteristics. Now we have two triangles on the family of characteristics, having a common vertex at the given Q-point (labelled "B" in Fig. 3.14).

Measuring the sides of these triangles with the help of the reference grid twice, we obtain ΔI_p, ΔV_g and ΔV_p as the arithmetic means of the two respective measurements

$$\Delta I_p = \frac{(I_p'' - I_p') + (I_p' - I_p)}{2} \tag{3-52}$$

$$\Delta V_g = \frac{(|V_g''| - |V_g'|) + (|V_g'| - |V_g|)}{2} \tag{3-53}$$

$$\Delta V_p = \frac{(V_p'' - V_p') + (V_p' - V_p)}{2} \tag{3-54}$$

The parameters can be found by using the following formulas

$$g_m = \Delta I_p / \Delta V_g \tag{3-55}$$

$$R_p = \Delta V_p / \Delta I_p \tag{3-56}$$

$$\mu = \Delta V_p / \Delta V_g \tag{3-57}$$

If two characteristics are available, we may limit ourselves to one triangle. Also, the transconductance of a triode can be determined from a single plate-grid characteristic, and the dynamic plate

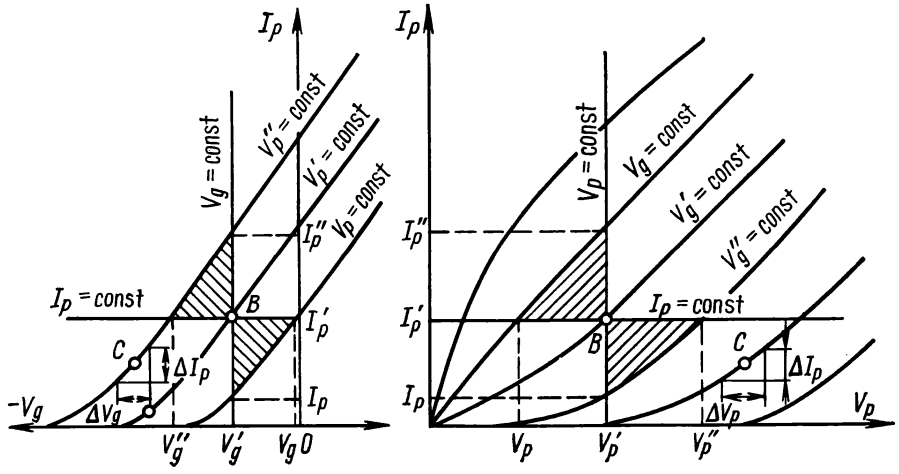


Fig. 3-14. Determining tube parameters from families of static characteristics

resistance from a single plate characteristic. This is carried out by finding small changes near the Q-point (C in Fig. 3-14). The amplification factor is determined as the product of g_m and R_p.

The applications of vacuum triodes are many and diverse. Accordingly, the tube parameters and the location of the Q-point should meet diverse requirements. For example, in some amplifier circuits, g_m must be a maximum; in other circuits, the same may be true of the dynamic plate resistance.

Families of respective characteristics also provide a qualitative insight into the tube parameters. For example, we can easily see that the slope of the plate-grid characteristic near the cut-off voltage is always smaller than at grid voltages close to zero. It is also an easy matter to find that at $V_g > 0$ the gradually sloping porton of

the plate characteristic corresponds to a higher, and the steeper portion, to a lower dynamic plate resistance.

Interelectrode Capacitances. When deriving the expression for the composite controlling voltage (Sec. 3-1), we mentioned two interelectrode capacitances, namely C_{gk} and C_{pk}. In addition to these two capacitances, the vacuum triode has a third capacitance, C_{pg}, between the grid and the plate.

When the cathode is hot (that is, emitting), the interelectrode capacitances vary due to the effect of the space charge. The potential distribution curve is most nonlinear between the grid and the cathode where a high-density space charge exists. The values of the capacitances, especially C_{gk}, are also affected by the thermal expansion of electrodes and their leads, variations in the permittivity of insulators, and some other factors. In amplifier triodes, the "hot" value of C_{gk} is greater than its "cold" value by 30 to 50%. The "hot" values of C_{pg} and C_{pk} differ but slightly from "cold" ones, because the potential distribution in the grid-plate space may well be deemed linear.

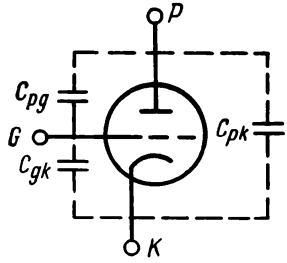


Fig. 3-15. Interelectrode capacitances of a triode

The capacitances C_{gk}, C_{pk} and C_{pg} in a triode shunt the respective interelectrode spaces (Fig. 3-15). Accordingly, they are called the *grid-to-cathode* (or *input*) *capacitance*, the *plate-to-cathode* (or *output*) *capacitance*, and the *plate-to-grid* (or *transfer*) *capacitance*, respectively.

The values of C_{gk}, C_{pk} and C_{pg} specified in data sheets represent the sums of the capacitances between the electrodes and their leads.

At low frequencies, the reactance of these capacitances is very high and their shunting effect is negligible. At high frequencies, they are comparable with the input and output resistances of a tube. At frequencies of about tens and hundreds of megahertz, they disturb the normal operation of a triode. In fact, they may do harm already at lower frequencies. For example, in a triode operating as an amplifier, the plate-to-grid capacitance, C_{pg}, responsible for an undesirable conductive path between the output and the input of the tube, may cause the amplifier to jump into oscillation.

As a rule, the interelectrode capacitances vary from 2 to 15 pF, but in high-power transmitting tubes they exceed 30 pF.

Plate Dissipation. As with the vacuum diode, the plate dissipation of the triode is defined by Eq. (2-31)

$$P_p = I_p V_p$$

Maximum safe values of plate dissipation, $P_{p,\max}$, are usually specified in data sheets. A curve representing the maximum plate

dissipation plotted on a family of plate characteristics (Fig. 3-16) limits the values of triode currents and voltages at which the tube will be run safely.

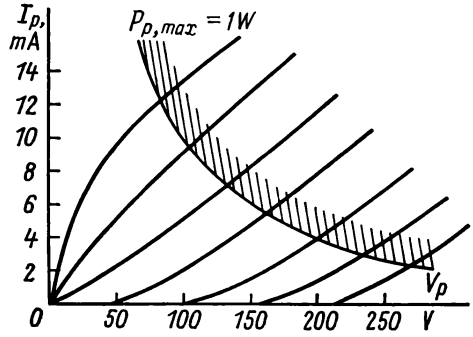


Fig. 3-16. Maximum power curve

3-4. OPERATION OF THE VACUUM TRIODE UNDER LOAD

Up to now, the discussion of the triode operation has been based on the assumption that a change in the voltage at one of the electrodes has no effect on the voltages at the other electrodes and affects only tube currents. Such an operation might be achieved in practice, if all supply circuits were independent and the voltage sources and connecting wires had negligible resistances.

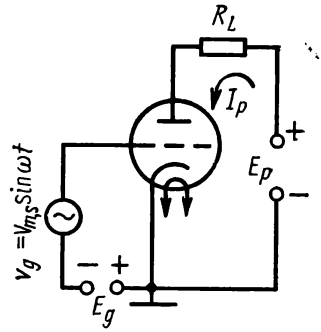


Fig. 3-17. Load connected in the plate circuit of a triode

Load in the Plate Circuit of the Triode. In practical circuits, the various circuits of a triode contain resistors, inductors, resonant circuits and some other components. These make up what is called the load. Figure 3-17 shows a triode amplifier with a load resistor in the plate circuit. For simplicity, the filament circuit is shown by two arrows. The grid circuit contains a direct voltage source, E_g, whose positive terminal is grounded, and also an oscillator generating the signal, v_g, to be amplified. The plate circuit contains a direct plate voltage source, E_p, and a load resistor, R_L, connected in series.

Let us discuss the operation of this circuit. To begin with, we assume that the plate load resistance is zero, $R_L = 0$, and, as a consequence, the plate voltage with respect to the cathode is equal to E_p. The direct negative voltage, E_g, applied between the grid and the cathode and called the *bias voltage* (or, simply, *bias*) positions

the Q-point on the characteristic shown in Fig. 3-18*a*. If, in addition to the bias voltage, a sinusoidal voltage, $v_g = V_b \sin \omega t$, is applied to the grid, the Q-point will be moving in step with the sinusoidal voltage along the characteristic (E_p = const) between points B and C where the total grid voltage is $E_g - V_b$ and $E_g + V_b$, respectively. The line along which the Q-point moves in step with the instantaneous grid voltage is termed the *path of operation*. In the plot of Fig. 3-18*a*, it is represented by a straight-line segment, BC, parallel to the y-axis, because, as v_g varies, the plate voltage E_p remains constant. The plate current varies sinusoidally with v_g as shown by the dashed line in Fig. 3-18*b*.

Thus, at $R_L = 0$, the alternating voltage applied to the grid brings about in-phase variations in plate current; the plate voltage remains constant and equal to the voltage supplied by the plate battery.

Now let us discuss the case where the load resistance is non-zero. The reasoning is the same as in the case of the vacuum diode (see Sec. 2-4). According to Kirchhoff's voltage law, we may write for the loop carrying the plate current (Fig. 3-17)

$$E_p = V_L + V_p \tag{3-58}$$

where $V_L = I_p R$ is the voltage drop across the load resistor, and V_p is the voltage drop across the plate-cathode space of the tube.

We also assume that the negative grid bias, E_g, remains the same. Under these conditions, the Q-point moves to another plate-grid characteristic corresponding to $V_p = E_p - V_L$ (point Q' in Fig. 3-18*a*). The voltage drop across the load resistor, the new value of the quiescent plate current, I_{p0}, and the position of the operating point, Q', on the family of characteristics may be found by Eq. (3-58).

The Load Line. Let us re-write Eq. (3-58)

$$I_p = E_p/R_L - V_p/R_L \tag{3-59}$$

As is seen, under load a plot of plate current against plate voltage yields a straight line with a slope $-1/R_L$. The first term on the right-hand side of Eq. (3-59) is a constant value because E_p = const and R_L = const. The straight line described by Eq. (3-59) can easily be constructed by finding the intercepts which it makes on the current and voltage axes: at $I_p = 0$, $V_p = E_p$, and, at $V_p = 0$, $I_p = E_p/R_L$. The "minus" sign of the slope indicates that the plate current represented by the line increases with decreasing plate voltage. The line shown in Fig. 3-18*c* is called the *load line*. It determines the plate current at constant values of E_p and R_L. As will be recalled the static plate characteristics of the triode also determine I_p, but they do so at constant V_g and $R_L = 0$.

Thus, the intersection of the load line with the respective plate characteristic locates the Q-point and defines the plate current of

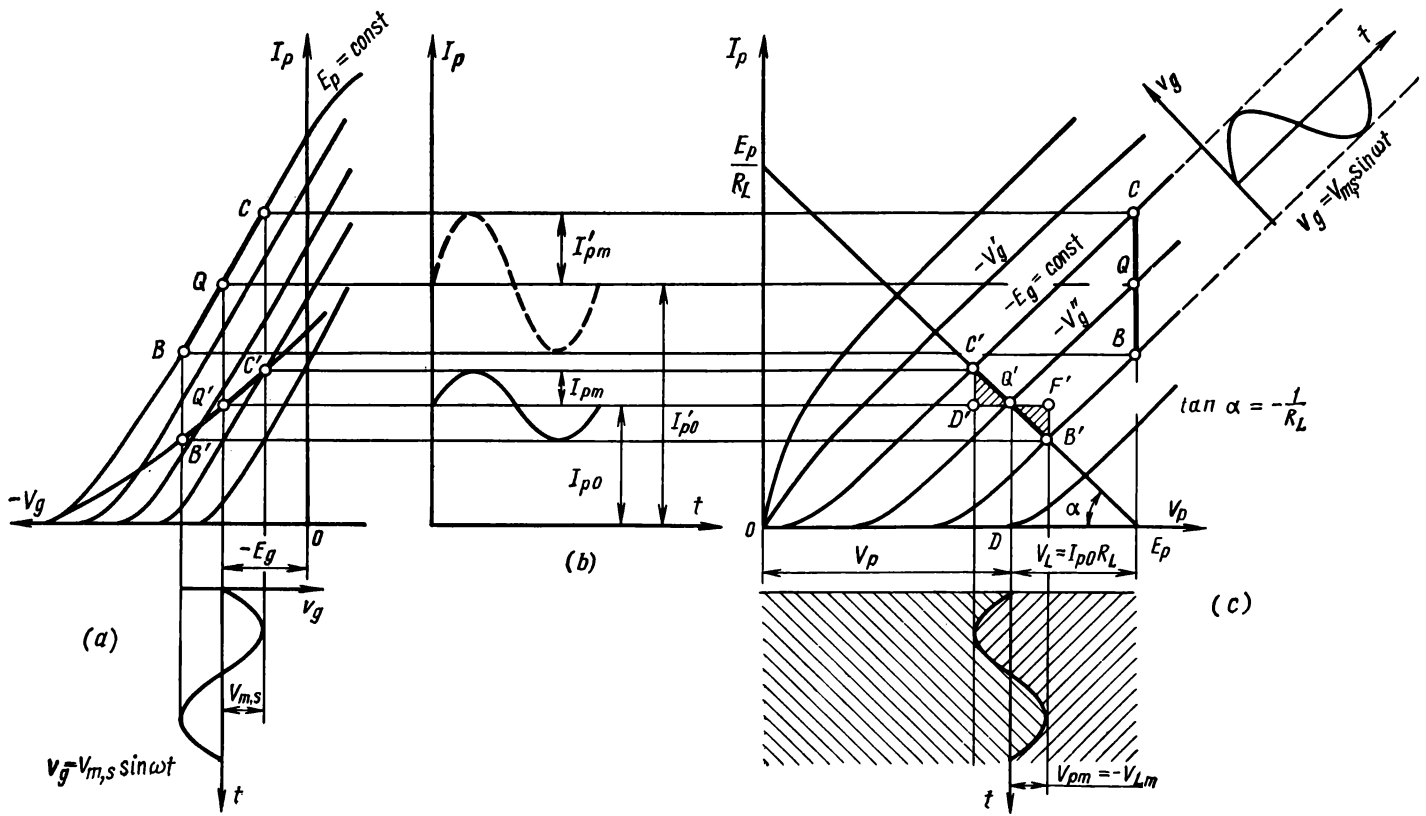


Fig. 3-18. Operation of a triode under load

(*a*) plotting the load line on a family of plate-grid characteristics; (*b*) plate current waveforms; (*c*) plotting the load line on a family of plate characteristics

the triode containing the load R_L in its plate circuit. In effect, what we have is a graphical solution of Eq. (3-59) simultaneously with the equation of the static plate characteristic, $I_p = f_2(V_p)$.

So, with no alternating (signal) voltage applied to the grid and at the given values of negative bias, E_g, plate supply voltage, E_p, and the plate load, R_L, the Q-point of the triode is located by the intersection of the load line with the static plate characteristic corresponding to $-E_g = \text{const}$ (at Q' in Fig. 3-18*c*). A projection of this point on the y-axis gives the quiescent plate current, I_{p0}. A line drawn through point Q' at right angles to the x-axis divides the segment OE_p into two parts. As is seen from the triangle $Q'DE_p$, the segment DE_p is equal to $I_{p0}R_L$, that is, the voltage drop across the load resistor R_L. According to Eq. (3-58), the remaining part of OE_p is equal to the voltage drop, V_p, across the plate-cathode space.

Now, if we apply a sinusoidal voltage, $v_g = V_{m,s} \sin \omega t$, to the grid, the Q-point (Fig. 3-18*c*) will be moving along the load line in step with this voltage. The limits of its path, B' and C', will correspond to the maximum and minimum negative grid voltages (in absolute values), $-V_g'' = -E_g - V_{m,s}$ and $-V_g' = -E_g + V_{m,s}$, respectively.

The path of the Q-point and, as a consequence, the load line may be constructed on the family of plate-grid characteristics by transferring points B' and C' to the I_p, V_g system of coordinates.

Referring to Fig. 3-18*a*, the load line plotted on the family of plate-grid characteristics is more gradually sloping than the static characteristics. So, the amplitude of the alternating component of the plate current, I_{pm} (shown by the solid sinusoid in Fig. 3-18*b*), becomes smaller than when $R_L = 0$ (the dashed line).

When the grid becomes less negative, the plate current rises and, as a consequence, the voltage across the load resistor, V_L, increases. Under the circumstances, according to Eq. (3-58), the plate voltage V_p falls and prevents the plate current caused by the variation in V_p from rising. In contrast, when the grid becomes more negative, there is a decrease in I_p and V_L and an increase in V_p, and this prevents the plate current from going down.

Thus, when the vacuum triode operates under load, the plate current and the load voltage vary simultaneously and in phase with the grid voltage. The voltage across the plate-cathode space varies in anti-phase with the grid voltage, thereby setting a limit for variations in plate current (Fig. 3-18*c*).

This is known as *plate reaction*, because variations in plate voltage oppose, as it were, variations in grid voltage in controlling plate current. Plate reaction decreases with decreasing effect of V_p on the plate current, that is, with increasing dynamic plate resistance or amplification factor, μ.

The efficiency of plate voltage in controlling plate current characterizes the performance of triode as an amplifier, because a change in plate voltage, $| dV_p | = dI_p R_L$, is the signal voltage at the triode output, and the dynamic amplification factor of a triode is

$$K = dV_p/dV_g \tag{3-60}$$

where dV_g is the signal voltage at the grid, that is, at the input of the amplifier.

Thus, the efficiency of a tube amplifier increases with increasing dynamic plate resistance, R_p, or amplification factor, μ.

It should be remembered that when a triode operates as an amplifier, the output signal is stepped up in both voltage and power as compared with the input signal by drawing energy from the tube supply sources. This energy is converted into the energy of the signal by the processes taking place in the tube itself. As is seen, the triode acts as a converter of electric energy.

Construction of the Load Line. With E_p and R_L specified in advance, the load line may be constructed on plate characteristics by using the slope (Fig. 3-18*c*)

$$\tan \alpha = -(p/q)\,(1/R_L) \tag{3-61}$$

Here, p and q are the scaling factors for the y- and x-axis, respectively.

As a rule, plate current is plotted as ordinate in milliamperes and plate voltage as abscissa, in volts. Then for $\tan \alpha$ to be determined correctly, plate load should be expressed in kilohms.

In engineering practice, the Q-point (see Fig. 3-18*c*) and the plate voltage, E_p, are generally specified in advance. Then, to construct the load line, one locates two points on the axes of coordinates and finds the required value of R_L as the quotient of E_p and the current shown as the intercept made on the y-axis by the load line.

If the Q-point and the load resistance R_L are specified in advance, the Q-point should be projected on the voltage axis. The intercept thus made on the voltage axis will give the potential difference, V_p, between the plate and the cathode. The necessary plate supply voltage, E_p, is determined by adding together V_p and the product of I_{p0} and R_L.

A family of load lines for several fixed values of R_L and the given Q-point is a series of straight lines passing through that Q-point (Fig. 3-19). The higher R_L, the more gradual is the load line. The limiting cases are a straight line parallel to the y-axis (at $R_L = 0$) and a straight line parallel to the x-axis (with the plate circuit open-circuited, $R_L = \infty$).

Dynamic Transconductance. As already stated, the load line plotted on a family of plate-grid characteristics is more gradually sloping than the static characteristics (Fig. 3-18*a*). When a triode

operates under load, the relationship between its plate current and grid voltage is given by the *dynamic transconductance* (or *dynamic mutual conductance*)

$$g'_m = dI_p/dV_g \quad R_L \text{ and } E_p \text{ remaining constant} \qquad (3\text{-}62)$$

Here, dI_p is the rate of change of plate current with grid voltage, when the plate circuit contains load. In contrast to the static (no-load) conditions (V_p = const), here V_p varies with V_g. To establish the relation between g_m and g'_m, it is necessary to find dV_p, that is, the change in plate voltage caused by a change in grid voltage, dV_g,

$$dV_p = -R_L dI_p \qquad (3\text{-}63)$$

Substituting Eq. (3-63) into Eq. (3-40), we get

$$dI_p = g_m dV_g - (R_L/R_p)\, dI_p \qquad (3\text{-}64)$$

Hence, recalling Eq. (3-62),

$$g'_m = dI_p/dV_g = g_m/(1 + R_L/R_p) \qquad (3\text{-}65)$$

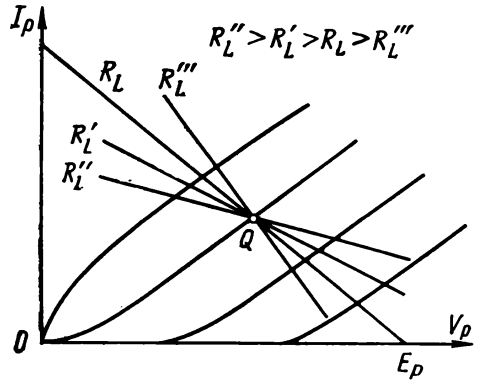


Fig. 3-19. Family of load lines for different loads

From the above expression it follows that g'_m is equal to g_m only under static (no-load) conditions ($R_L = 0$). When $R_L > 0$ (that is, under load), g'_m is always smaller than g_m. When $R_L = \infty$ (the plate circuit is open-circuited), the load line coincides with the axis of abscissa.

Some authors relate g'_m to g_m in a different way which can easily be derived from Eq. (3-65)

$$g'_m = \mu/(R_p + R_L) \qquad (3\text{-}66)$$

Dynamic Voltage Amplification Factor. Another important parameter characterizing the operation of a tube under load is the *dynamic voltage amplification factor*, symbolized k_v, which is defined as the ratio of a small change in plate voltage, dV_p, to the small change in grid voltage, dV_g, that caused it

$$k_v = \left| \frac{dV_p}{dV_g} \right| \quad R_L \text{ and } E_p \text{ remaining constant} \qquad (3\text{-}67)$$

Equation (3-62) suggests that

$$dI_p = g'_m dV_g \qquad (3\text{-}68)$$

Substituting Eq. (3-68) into Eq. (3-63), we get

$$dV_p = -g'_m R_L dV_g \qquad (3\text{-}69)$$

Using Eq. (3-69) and recalling Eqs. (3-66) and (3-67), we may establish the relation between the dynamic voltage amplification factor and the static (no-load) amplification factor

$$k_v = \mu R_L/(R_p + R_L) = \mu/(1 + R_p/R_L) \qquad (3\text{-}70)$$

As is seen, at no-load ($R_L = 0$), the dynamic voltage amplification factor is zero ($k_v = 0$). When $R_L = \infty$ (the plate circuit is open-circuited), $k_v = \mu$.

The dynamic transconductance, g'_m, and the voltage amplification factor, k_v, can be determined with the help of a load line plotted in the plate-grid coordinates. For this purpose, we construct a triangle near the Q-point, and find g'_m. Then, knowing R_L, we can find k_v by Eqs. (3-67) and (3-68).

The load line plotted in the plate coordinates makes it possible to determine k_v. To do this, the intercept made on the load line by two characteristics adjacent to the Q-point should be projected on the V_p-axis (for example, the projection of points B' and C' in Fig. 3-18*c*). The projection of this intercept is equal to $2V_{Lm}$ if the alternating grid voltage swings between the amplitudes V'_g and V''_g. The dynamic voltage amplification factor, k_v, is equal to the ratio of the intercept as projected on the x-axis to the voltage difference $| V''_g | - | V'_g |$.

Using the load line plotted in the plate coordinates, we may determine the power dissipated by the alternating component of the plate current across the load resistor

$$P_L = (1/2) I_{pm} V_{pm} \qquad (3\text{-}71)$$

Referring to Fig. 3-18*c*, $P_L = (1/2)\ Q'D' \times C'D'$, which shows that the power dissipated across R_L is determined by the surface of the shaded triangle constructed near the Q-point.

3-5. CONSTRUCTION AND TYPES OF VACUUM TRIODES

According to their designated function, vacuum triodes differ not only in main parameters and characteristics, but also in construction. Figure 3-20 shows a sketch of a miniature vacuum triode.

The electrodes of vacuum triodes also markedly differ in design. Cathodes, grids and plates may be given the shape of a parallelepiped with end faces removed, a cylinder, or an elliptical cylinder. As a rule, the cathode, filamentary or indirectly heated, is of the oxide-coated type.

In low-power tubes, the cathodes are made from nickel which has a high melting temperature. The grids (fabricated from nickel, its molybdenum alloy or tungsten) are made of a fine wire wound

on frames (Fig. 3-21*a*). Grounded-grid vacuum triodes use grids of the rod construction (Fig. 3-21*b*).

Vacuum triodes intended for a.f. preamplification operate as voltage amplifiers. For a greater dynamic voltage amplification factor, they should have a high μ. At a high μ, the cut-off voltage, $V_{g,c\text{-}o}$, is low, but, since the signal to be amplified is small, this may be neglected. What is important is that, at $v_g < 0$, the region of the plate-grid characteristic corresponding to the applied signal should

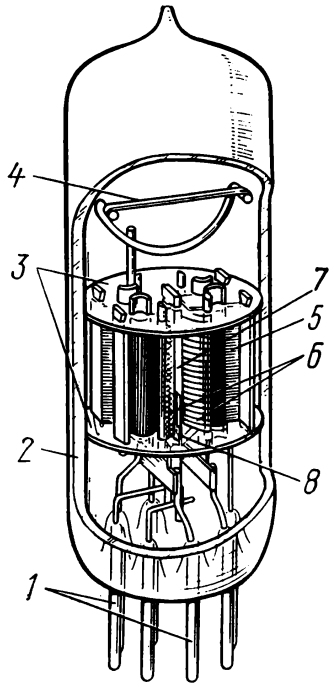


Fig. 3-20. Construction of a miniature double triode

1—pins; *2*—envelope; *3*—mica spacers; *4*—getter; *5*—plate; *6*—grid; *7*—cathode; *8*—heater

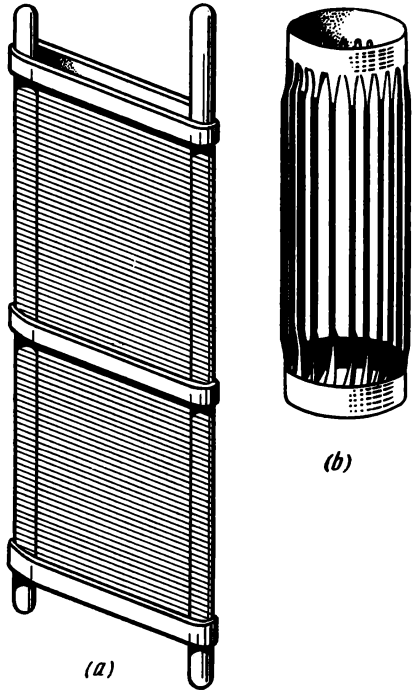


Fig. 3-21. Tube grids

(*a*) frame grid; (*b*) punched rod grid

be linear, wherever possible. Otherwise, the waveform of the amplified signal will be distorted.

Vacuum triodes used as power amplifiers should meet other requirements. These tubes handle strong signals, so the path of operation should be broad at negative values of V_g. Such characteristics are often called "left-hand" characteristics in contrast to "right-hand" characteristics where the linear path of operation is found largely in the region of positive values of V_g. A "left-hand" characteristic can be obtained only at a low μ; for triodes operating as power amplifiers,

μ is equal to several unities. The power of an alternating signal in the plate circuit of an amplifier tube is maximal when $R_L = R_p$. It is proportional to the product of μ and g_m, which is known as the *figure of merit of a tube*. So, for triodes used as power amplifiers, the transconductance, g_m, should be as high as possible.

Vacuum triodes designed for operation in r.f. and pulse amplifiers should have low interelectrode capacitances. For this purpose, smaller electrodes or special envelopes and leads are used.

Chapter Four

MULTIELECTRODE AND SPECIAL-PURPOSE TUBES

4-1. TETRODES AND PENTODES. PHYSICAL PROCESSES IN MULTIELECTRODE TUBES

Definition. A multielectrode tube is an electron tube containing a cathode, a plate and several grids associated with a single electron stream.

Disadvantages of Triodes. As has already been noted in the previous chapter, the transfer capacitance, C_{pg}, equal to several picofarads in low-power triodes and several tens of picofarads in high-power triodes, may deleteriously affect the operation of an amplifier triode. Refering to Fig. 4-1, C_{pg} introduces feedback from the output to input circuit of an amplifier stage. At radio frequencies, the reactance of this capacitance becomes low, and as an r.f. signal is applied from the output to the input, the amplifier may jump into oscillation.

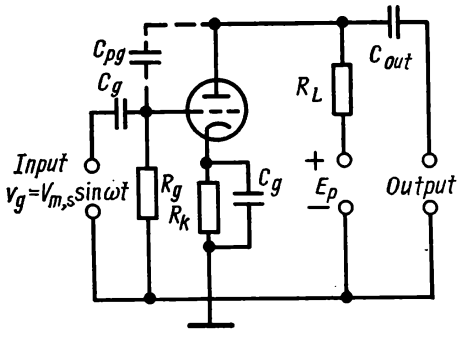


Fig. 4-1. Transfer capacitance in a triode amplifier

It is obvious that for stable operation of an amplifier at radio frequencies, the grid-plate capacitance should be reduced.

Another important drawback of the vacuum triode is its low no-load amplification factor, μ. Generally speaking, the amplification factor of the triode might be raised by reducing the spacing between the grid wires or by increasing the distance between the plate and the cathode. This, however, would give rise to other undesirable effects, because a rise in μ requires the operating plate voltage to be considerably increased. The point is that at $V_g < 0$, the cut-off potential of the triode is

$$| -V_{g,c\text{-}o} | = V_p/\mu \qquad (4\text{-}1)$$

If, for example, we needed, say, $\mu = 500$ and $V_{g,c-o} = -4$ V, the necessary plate voltage would be 2000 V. Such a high V_p would require more powerful supply sources and circuits and, in addition, a special tube design.

An acute need for a triode free from both disadvantages arose in the twenties when radio ventured into H. F. and VHF/UHF bands. To improve the performance of the triode, it was suggested that a fourth electrode — a *screen grid*—should be added to the usual plate, grid and cathode. This gave birth to a tetrode.

The Vacuum Tetrode. This four-electrode tube has the screen grid placed between the control grid and the plate. Physically, the screen grid (which is usually a wire helix fastened to supports) is similar to the control grid in the triode (Fig. 3-21*a*), but the spacing between its wires is smaller. Transmitting tetrodes sometimes use screen grids made of gauze. The screen grid is maintained at a positive potential usually equal to 60-100% of the plate voltage. At the signal frequency, the screen grid is returned to ground by a high-value bypass capacitor (Fig. 4-2), so the resultant capacitance between the control grid and the plate is considerably reduced owing to the shielding effect of the second grid. In tetrodes, the capacitance between the plate and control grid, C_{pg1}, is equal to a few hundredths of a picofarad, thus being about a hundredth of C_{pg} in the triode.

The dense screen grid markedly reduces the penetration of lines of force from the plate into the space-charge region at the cathode. Accordingly, the penetration factor of the tetrode is smaller, and the dynamic plate resistance, R_p, and the no-load amplification factor, μ, are higher than those in the vacuum triode.

As compared with the triode, the tetrode has another distinction. Here, not only the plate, but also the screen grid is held at a positive potential. So, the electron stream from the cathode divides into two streams always (even if $V_{g1} < 0$): one going towards the plate and the other towards the screen grid. When the control grid is positive, some electrons also fall on the control grid.

Tetrode Characteristics. The tetrode characteristics, $I_p = f(V_p)$ and $I_{g2} = f_1(V_p)$, derived analytically are shown in Fig. 4-3 by the dashed lines. The characteristics of an actual tetrode, however, have a different shape (solid lines in Fig. 4-3). This can be explained as follows. At $V_p = 0$, all of the electrons which have passed through the control grid fall on the screen grid, so $I_k = I_{g2}$ and $I_p = 0$. At a low positive V_p, the plate accepts a small proportion of electrons which have passed through the screen grid. Most of these electrons return to the screen grid wires, thus making up a space charge between the screen grid and the plate (region *1* in the figure, which represents electron flyback). As V_p rises, plate current increases, too, and I_{g2} decreases because the plate field (not being screened by the grid) directly affects the electrons between the screen grid and the

plate. At $V_p \geqslant 20$ V, the electrons reaching the plate have enough energy to knock out secondary electrons from its surface. Under the action of a higher voltage, V_{g2}, these electrons fall on the screen grid and contribute to I_{g2}. The plate current decreases because secondary electrons leaving the plate for the grid make up a current in the plate circuit, which is opposite to the current due to the primary electrons. This is represented by the drooping region on the $I_p = f(V_p)$ characteristic (region *2* in Fig. 4-3), corresponding to a negative a.c. resistance: plate current decreases with rising V_p. As the plate voltage grows further, I_p begins to rise again, and I_{g2}

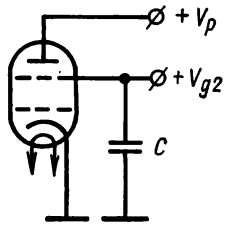


Fig. 4-2. Screen grid circuit

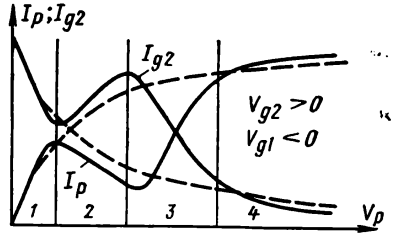


Fig. 4-3. Plate characteristics of the tetrode (dashed lines — rated characteristics, solid lines — actual characteristics)

decreases (region *3* in Fig. 4-3) because still more secondary electrons fly back to the plate. This is known as the *dynatron effect.*

Further on, the ideal characteristics run gradually (region *4* in Fig. 4-3): as the plate voltage rises, the plate current slowly increases and the grid current, I_{g2}, decreases in proportion, which corresponds to electron interception. The gradual region on the ideal characteristic $I_p = f(V_p)$ does not imply that the tube has reached saturation. The negative space charge at the cathode surface is sufficiently high (if no considerable positive voltage is applied to the control grid). A small rise in the plate current at a significant change in V_p in this region can be ascribed to a minor effect which the plate voltage has on the current partition ratio between the screen grid and the plate.

In this part of the actual tetrode characteristics, I_p increases and I_{g2} decreases because of the secondary emission from the screen grid; secondary electrons move from the screen grid to the more positive plate.

As is seen from Fig. 4-3, the dynatron effect accounts for the difference between the actual and ideal tetrode characteristics; due to this effect, the tetrode cannot be used at low plate voltages (corresponding to regions *2* and *3*). We have to choose plate load and source voltage E_p such that the load characteristic would inter-

sect the static plate characteristics in region *4*. Because of this tetrodes have not found wide application. Various measures have been suggested to eliminate the dynatron effect in tetrodes. Their main idea is to create a potential barrier between the control grid and the plate for secondary electrons emitted by these electrodes. In other words, a low potential region, V_{min}, should be established between the screen grid and the plate. This can be achieved either

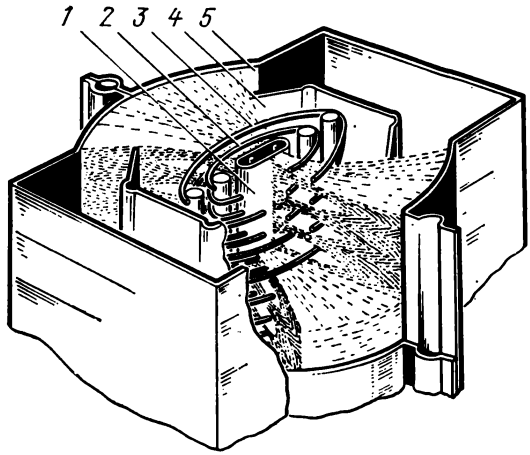


Fig. 4-4. Beam-power tetrode
1—cathode; *2*—control grid; *3*—screen grid; *4*—beam-confining plate; *5*—plate

by using a space charge or by placing an additional (third) grid between the screen grid and the plate. The tubes utilizing the first method are called *beam-power tetrodes*; those utilizing the second method are known as *pentodes*.

The Beam-Power Tetrode. A sketch of a beam-power (or, simply, beam) tetrode is shown in Fig. 4-4. The cathode and grids are of elliptical configuration, and the plate is cylindrical in its operating part. Owing to this, the distance between the screen grid and the plate is increased and secondary electrons cannot move from the plate to the grid so easily because the screen grid field has a smaller effect near the plate.

Between the screen grid and the plate are placed two beam-confining (or beam-forming) plates connected to the cathode within the tube. Because of these plates, electrons travel from the cathode to the plate within a limited space in sheetlike beams which have given the name to the tube.

The wires in the control and screen grids are spaced equal distances apart and aligned so that the screen-grid wires are in the shadow of the control-grid wires as seen from the cathode. Because of this construction, the beamlike electron stream is divided into several

beams in the vertical plane. Electron beams increase the density of the negative space charge, and a potential barrier appears in the space between the screen grid and the plate, which prevents secondary electrons from travelling from the plate to the screen grid at $V_p < V_{g2}$, and from the screen grid to the plate at $V_p > V_{g2}$.

Figure 4-5*a* shows potential distribution curves for the beam and the conventional tetrode when the plate is at a low potential and the tube is subject to the dynatron effect (region *2* in Fig. 4-3).

Between the cathode and the screen grid, these curves practically coincide; between the screen grid and the plate, a low potential

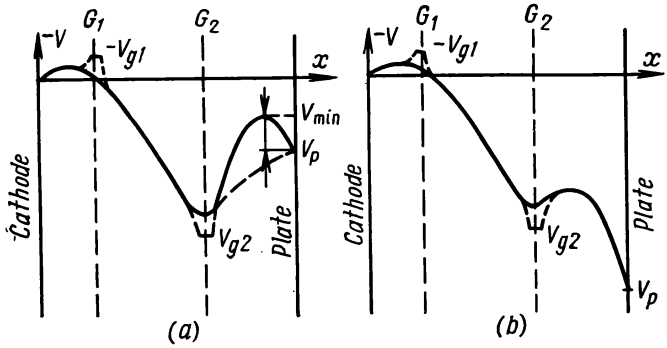


Fig. 4-5. Potential distribution
(*a*) in the beam (the solid line) and conventional (the dashed line) tetrode in the presence of the dynatron effect; (*b*) in the beam tetrode under nominal conditions

barrier exists in the beam tetrode, to overcome which the initial velocities of secondary electrons are insufficient. In the beam tetrode, the dynatron effect is suppressed, if the potential at the V_{min}-point is 15 to 20 V below that at the plate. The value of the potential barrier and the distance between the V_{min}-point and the plate depend on both the relative values of V_p and V_{g2} and the density of the space charge between the screen grid and the plate which is determined by the control grid potential. With a high-density electron stream, the potential between the screen grid and the plate may decrease down to zero. At the point where $V_{min} = 0$, electrons have zero velocities ($v = 0$), so this region may be regarded as a cathode emitting electrons having zero velocities. Because of this, the interelectrode region where $V_{min} = 0$ is called the *virtual cathode.*

The potential distribution in the beam power tetrode at nominal V_p and V_{g2} is shown in Fig. 4-5b.

The Pentode. As already noted, the other method of getting rid of the dynatron effect leads to the *pentode.* The vacuum pentode, as its name implies, is a five-element tube. It is like a screen-grid tetrode with another grid added between the screen grid and the plate. The additional grid is called the *suppressor grid.* It is a coarse

grid connected to the cathode and, as consequence, operated at zero potential. The potential barrier between the plate and the screen grid which does not allow secondary electrons to emerge from the plate owing to the dynatron effect is established by the suppressor grid field (curve *a* in Fig. 4-6). The potential distribution in the pentode at nominal V_p and V_{g2} is given in Fig. 4-6 by curve *b*.

The suppressor grid not only eliminates the dynatron effect, but also markedly improves the no-load parameters of the tube as compared with the triode and tetrode. Pentodes have a very high amplification factor running up to several thousand, a great dynamic plate resistance and a low transfer capacitance. Pentodes specially designed for r.f. operation have a number of additional features to improve these parameters.

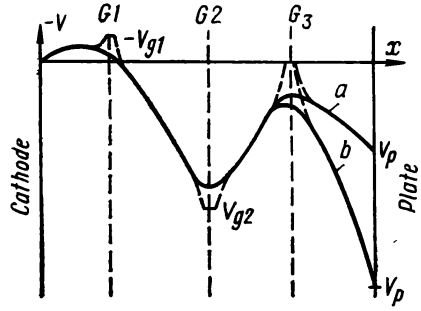


Fig. 4-6. Potential distribution in the pentode
(*a*) under conditions of the dynatron effect; (*b*) under nominal conditions

The Composite Controlling Voltage in Multielectrode Tubes. To compare the effect which the plate and grid voltages of the triode have on the space charge near the cathode, we replaced the triode by an equivalent diode (see Sec. 3-1) with the plate placed in the same plane with the grid. In the equivalent diode, the plate potential has the same effect on the space charge at the cathode

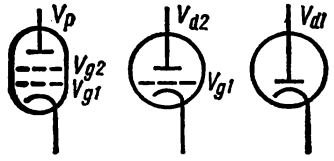


Fig. 4-7. Tetrode replaced by an equivalent diode

as the actual triode plate and grid potentials taken together. Applying this method, we can replace the tetrode by a certain equivalent triode whose plate is placed in the same plane with the screen grid (Fig. 4-7). Now, the plate of the equivalent triode should be at the following potential

$$V_{d2} = (\text{approx.})\ V_{g2} + D_2 V_p \qquad (4\text{-}2)$$

Here, D_2 is the *penetration of the screen grid* which shows how much the lines of force extend from the plate into the space between the control grid and the screen grid.

Next, the equivalent triode may be replaced by an equivalent diode. For this purpose, we replace the grid by the virtual plate. Then, on the basis of Eq. (3-8) for the composite controlling voltage in the triode and recalling Eq. (4-2), we may write an expression for the composite controlling voltage in the tetrode

$$\begin{aligned} V_{d1} &= (\text{approx.})\ V_{g1} + D_1 (V_{g2} + D_2 V_p) \\ &= (\text{approx.})\ V_{g1} + D_1 V_{g2} + D_1 D_2 V_p \end{aligned} \tag{4-3}$$

Here, V_{d1} is the composite controlling voltage at the plate of the equivalent diode placed in the same plane with the control grid of the tetrode; D_1 is the *penetration of the control grid* showing how much the field established by the electrodes behind this grid extends towards the cathode surface; $D = D_1 D_2$ is the *penetration of the tube*, which shows how much the lines of force extend from the plate through the screen grid and the control grid into the space-charge region at the cathode, and how strong the plate and control grid potentials are in affecting the space charge near the cathode.

The penetration of the control grid, D_1, is usually equal to several hundredths; the penetration of the screen grid, D_2, whose wires are spaced closer apart, is even smaller. Since $D_1 D_2 \ll 1$, the last term in Eq. (4-3) may be neglected. Then

$$V_{d1} = (\text{approx.})\ V_{g1} + D_1 V_{g2} \tag{4-4}$$

In a like manner, we may find the composite controlling voltage in the pentode. The voltage at the plate of the equivalent tetrode should be equal to the composite controlling voltage in the plane of the suppressor grid of the pentode

$$V_{d3} = V_{g3} + D_3 V_p + D'_3 V_{d2} \tag{4-5}$$

The second and the third terms on the right-hand side of the equation describe the effect that the plate field and the composite controlling voltage have on the field in the plane of the suppressor grid. Here, D_3 is the penetration of the suppressor grid, which shows how much the lines of force extend from the plate through this grid to the cathode; D'_3 is the reverse penetration of the suppressor grid, which shows how much the field established by the composite controlling voltage in the plane of the screen grid extends through the suppressor grid wires into the space between the plate and the suppressor grid.

The reverse penetration factors, D'_1 and D'_2, have not been taken into consideration before because the wires of the control grid and the screen grid are spaced considerably closer apart than those of the suppressor grid. In the pentode where the suppressor grid wires are spaced wide apart, D'_3 must sometimes be taken into account.

Using Eq. (4-5) and recalling that V_{g3} is usually zero, we find the voltage at the plate of the equivalent triode, which is equal to the composite controlling voltage in the plane of the screen grid

$$V_{d2} = V_{g2} + D_2 V_{d3} = (V_{g2} + D_2 D_3 V_p)/(1 - D_2 D_3') \quad (4\text{-}6)$$

Using the above expression, we can easily find the plate voltage of the equivalent diode

$$V_{d1} = V_{g1} + D_1 V_{d2} = V_{g1} + (D_1 V_{g2} + D_1 D_2 D_3 V_p)/(1 - D_2 D_3') \quad (4\text{-}7)$$

Usually, $D_1 < 0.1$, $D_2 < 0.1$, $D_3 < 0.5$ and $D_3' < 0.5$. So, $D_1 D_2 D_3 \ll \ll 1$ and $D_2 D_3' \ll 1$. Accordingly, in most practical cases, we may usé a simplified expression for the pentode, similar to Eq. (4-4) for the tetrode

$$V_{d1} = (\text{approx.})\ V_{g1} + D_1 V_{g2}$$

Owing to the screening effect of the grids in the tetrode and pentode, the effect of the plate voltage on the space charge near the cathode is negligibly small.

Current Partition in Multielectrode Tubes. In the tetrode and pentode, the screen grid maintained at a high positive potential divides the electrons emitted by the cathode into two streams: one moving towards the plate and the other towards the screen grid. In contrast to the triode which operates in some circuits with the grid drawing no current and where current partition takes place only at $V_g > 0$, in the tetrode and pentode the cathode current is the sum of two currents even at $V_{g1} < 0$ ($I_g = 0$)

$$I_k = I_p + I_{g2} \quad (4\text{-}8)$$

In some cases, positive voltages may be applied to the control and suppressor (in the pentode) grids. Then, for the tetrode

$$I_k = I_{g1} + I_{g2} + I_p \quad (4\text{-}9)$$

and for the pentode

$$I_k = I_{g1} + I_{g2} + I_{g3} + I_p \quad (4\text{-}10)$$

In the general case, the cathode current in multielectrode tubes is the sum of all grid currents and plate current (Fig. 4-8). Tetrodes, however, usually operate with the control grid held negative ($V_{g1} < 0$), and pentodes, at $V_{g1} < 0$ and $V_{g3} = 0$. Quantitatively, this is described in terms of the *current partition ratio*

$$k_{cp} = I_p/I_{g2} \quad (4\text{-}11)$$

which depends on the relative magnitudes of plate and control grid voltages (Fig. 4-9)

$$k_{cp} = f(V_p/V_{g2}) \quad (4\text{-}12)$$

Using the current partition ratio, I_p and I_{g2} may be written as fractions of the total cathode current

$$I_p = I_k k_{cp}/(k_{cp}+1) \quad (4\text{-}13)$$

$$I_{g2} = I_k/(k_{cp}+1) \quad (4\text{-}14)$$

As in the triode, current partition in multielectrode tubes takes two forms. When the plate is at a low potential as compared with V_{g2}, the paths of some electrons flying through the screen grid are bent by the resultant field between this grid and the plate, and the electrons return from the plate to the screen grid. This is electron

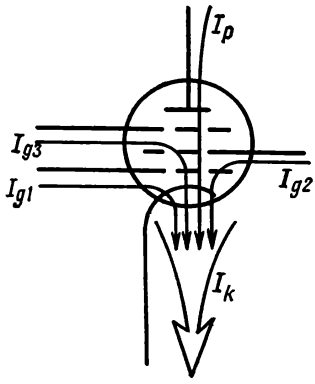


Fig. 4-8. Currents in the pentode

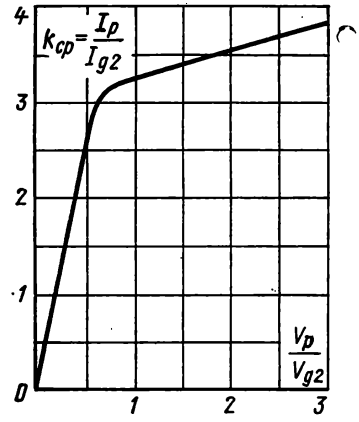


Fig. 4-9. Current partition ratio as a function of V_p/V_{g2}

flyback. In the tetrode, plate voltage directly affects the field in the screen grid—plate space, whereas in the pentode this effect is somewhat smaller due to the screening action of the coarse suppressor grid. So, in this case, even a small rise in plate voltage markedly affects the paths of the electrons which have passed through the screen grid. As V_p increases, the paths of a still greater proportion of electrons are straightened, I_p sharply rises and I_{g2} falls. As a consequence, the current partition ratio increases in proportion.

When electrons are directly intercepted by the screen grid, plate current is insignificantly affected by variations in plate voltage. Almost all electrons which have passed through the screen grid make for the plate. An insignificant rise in plate current with an increase in plate voltage can be attributed, firstly, to the fact that the current partition ratio increases as the number of electrons intercepted by the screen grid decreases, and secondly, to a very small rise in cathode current. The effect of plate voltage on the space charge near the cathode and cathode current is very small due to the screening action of the grids.

4-2. STATIC CHARACTERISTICS OF MULTIELECTRODE TUBES

The characteristics of various multielectrode tubes differ insignificantly. So, we shall only discuss the main characteristics of the pentode noting, where necessary, their difference from those of the beam power tetrode. In the pentode, the currents in the electrode leads are functions of all voltages. Since each characteristic represents the relation between each current and each voltage with the

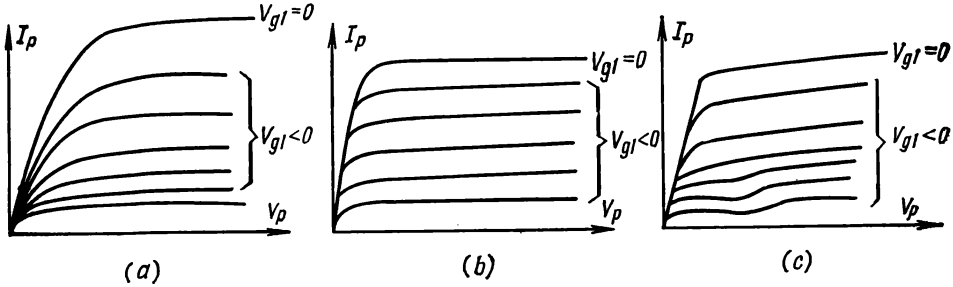


Fig. 4-10. Plate characteristics
(a) variable-μ pentode; (b) r.f. pentode; (c) beam-power tetrode

other three voltages maintained constant, as many as twenty characteristics may be derived for the pentode. The pentode, however, usually operates at $V_{g1} < 0$ ($I_{g1} = 0$), $V_{g3} = 0$ ($I_{g3} = 0$) and V_{g2} = constant. So we shall only consider I_p and I_{g2} as functions of V_{g1} and V_p.

Plate Characteristics. These constitute the main family of characteristics used when calculating circuits built around vacuum tubes. In data sheets for pentodes, these characteristics are specified for the nominal V_{g2} and $V_{g3} = 0$. Figure 4-10 shows the plate characteristics for two types of pentode and a beam power tetrode. Despite some differences, the plate characteristics of multielectrode tubes are rather typical. In any case, we can easily distinguish the operating region for the pentode—operation under conditions of electron interception. In this region, the $I_p = f(V_p)$ curves are very gradual; the dynamic plate resistance is high. It is also an easy matter to mark the non-operating region on the characteristics—operation under conditions of electron flyback. This region is represented by the sharply rising initial portion. Here, the dynamic plate resistance is low.

Let us discuss the difference between these curves in more detail. The transition region from electron flyback to electron interception varies from tube to tube. The range of plate voltages corresponding to this region depends on the density of the grids, the distance

between them, and also the distance between the screen grid (the suppressor grid in the pentode) and the plate.

The construction of the grids and the distances between the electrodes in tetrodes and pentodes depend on the purpose of each particular tube. As a rule, one seeks to reduce the current drawn by the screen grid in order to cut down power consumption. In addition, it is desirable that the transition from the steep initial portion (electron flyback) to the gradual portion (electron interception) should take place at the lowest possible plate voltage. This extends the region of the characteristics where operation is under conditions of electron interception.

For this purpose, the grids are designed and placed so that the paths of electrons moving to the plate are as close to straight lines as possible. In the beam power tetrode, these requirements are met quite well because the paths of electrons are bent mostly by the control grid; the screen grid whose wires are spaced the same distance apart as those of the control grid affects electron paths insignificantly.

In r.f. amplifier tubes, the screen grid is sought to be made as dense as possible. The distance between the grid and the other grids is also larger than in conventional tubes. This contradicts, however, the requirements to cut down the screen-grid current and to reduce the transit time of electrons, so a compromise must be found. In tubes whose screen grid wires are spaced close apart, the field in the plane of the screen grid is more uniform, and the transition from electron flyback to electron interception occurs more sharply.

To reduce the dissipation of electrons by the suppressor grid, its wires are spaced wide apart and it is placed as close to the plate as possible.

The slope of the characteristics in the operating region shows how much plate voltage affects the current partition ratio and the space charge near the cathode. In pentodes with dense screen grids and relatively dense suppressor grids, the $I_p = f(V_p)$ curves run almost parallel to the x-axis. To amplify high-amplitude signals, use is made of tubes having a large negative cutoff voltage. In pentodes, this requirement is met by using screen grids with wires spaced wide apart. As a consequence, in these tubes the plate characteristics are less gradual under conditions of electron interception.

At the beginning of electron interception, the dynatron effect produces an insignificant kink in the plate characteristics of some beam tetrodes. As a rule, this kink in the plate characteristic takes place at very negative V_{g1} when the density of the electron stream is low and insufficient to suppress the dynatron effect completely.

Plate-Grid Characteristics. These are shown in Fig. 4-11. According to Eq. (4-4), the composite controlling voltage in multielectrode

tubes does not practically depend on plate voltage

$$V_{c1} = \text{(approx.)}\ V_{g1} + D_1 V_{g2}$$

so cathode current

$$I_k = G(V_{g1} + D_1 V_{g2})^{3/2} \tag{4-15}$$

is equal to zero if

$$V_{g1} + D_1 V_{g2} = 0 \tag{4-16}$$

In Eq. (4-15), G is the constant of proportionality depending on the construction of a tube.

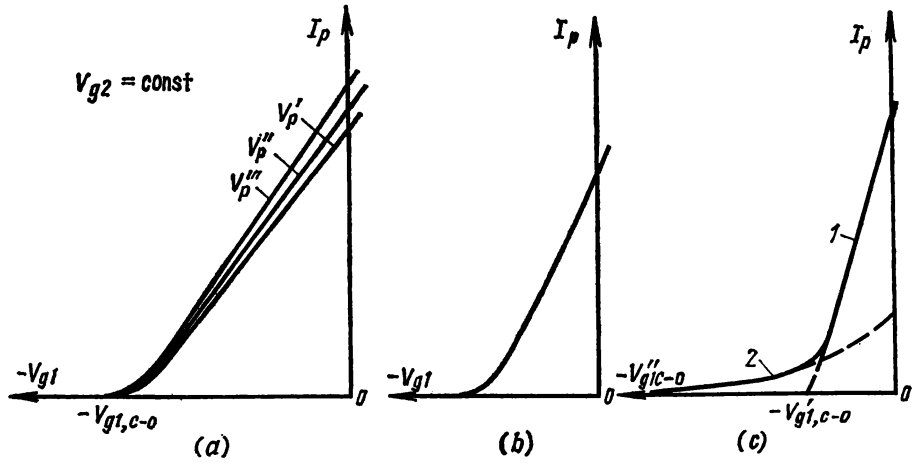


Fig. 4-11. Plate-grid characteristics
(*a*) family of characteristics; (*b*) sharp cut-off characteristic; (c) characteristic of a variable-μ tube

As is seen, cut-off voltage in multielectrode tubes ($I_k = 0$) does not depend on plate voltage and can be derived from the following expression

$$V_{g1,\,c\text{-}o} = -D_1 V_{g2} \tag{4-17}$$

Referring to Fig. 4-11*a*, all characteristics for V_{g2} = const and different V_p originate from the same point. As, under conditions of electron interception, plate voltage affects the space charge near the cathode very little, variations in V_p also have an insignificant effect on plate current; plate-grid characteristics constitute a series of very closely spaced curves originating from the same point in a fan-like manner. Because of this, data sheets either omit these characteristics altogether or specify only one characteristic plotted for nominal V_p and V_{g2}.

These characteristics markedly differ from tube to tube. Figure 4-11*b* shows what is known as a sharp cut-off characteristic plotted

for a tube having a high μ and, as a consequence, a low cut-off voltage, $V_{g1,c-o}$. Figure 4-11*c* shows a remote cut-off characteristic typical for variable-μ pentodes. In these tubes, a portion of the control grid (usually one or two turns) is made more coarse (Fig. 4-12). Such a pentode may be regarded as two tubes connected in parallel. The tube with the dense control grid has usual high g_m and μ. The other, corresponding to the portion of the control grid with the wires spaced wider apart, has a considerably lower amplification factor and a small transconductance. The steep region, *1*, of the characteristic (Fig. 4-11*c*) corresponds to the first tube; the gradual region, *2*, represents the second tube whose cut-off potential, $V''_{g1,c-o}$ exceeds $V'_{g1,c-o}$ due to the low μ. These tubes are used for automatic gain control in r.f. amplifiers.

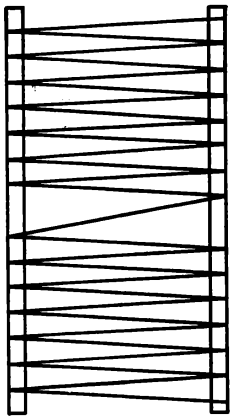


Fig. 4-12. Grid of variable-μ pentode

Control of Plate Current in the Pentode. In a pentode, plate current may be controlled not only by the control grid but also by the screen grid. So, in addition to the above static characteristics, the functions $I_p = \varphi(V_{g3})$ at $V_{g1} = \text{const}$ and $I_p = \varphi_1(V_{g1})$ at $V_{g3} = \text{const}$ (Fig. 4-13) are also of interest.

Let us first discuss the function $I_p = \varphi(V_{g3})$ at $V_{g1} = \text{const}$. When $-V_{g3} = -V_{g3,c-o}$, the resultant potential between the wires of the suppressor grid is negative. The retarding effect of this potential barrier is such that all electrons are forced to return to the screen grid. The suppressor-grid voltage, $-V_{g3,c-o}$, cuts off the tube for plate current, and the current drawn by the screen grid is equal to the cathode current, I_k. The control-grid voltage does not practically affect the composite controlling voltage in the plane of the suppressor grid, and $-V_{g3,c-o}$ remains the same for all values of $-V_{g1}$. As the suppressor grid becomes less negative, the potential barrier near it decreases. A still greater number of electrons make for the plate, I_p rises and I_{g2} falls. Thus, control of I_p by V_{g3} is due to the repartition of the electron stream between the screen grid and the plate. V_{g3} has a negligible effect on the space charge near the cathode, so the cathode current remains practically the same. As the control grid becomes more negative, there is a fall in cathode current and I_p which is a part of I_k.

The characteristics $I_p = \varphi_1(V_{g1})$ at $V_{g3} = \text{const}$ (Fig. 4-13*b*) have a different shape, because the plate current is controlled in a different way. Here, plate current is a function of the effect that the control grid has on the space charge near the cathode. As the control grid becomes less negative, a still greater proportion of electrons pass through the control grid. This increases not only the

cathode current, but also I_p and I_{g2}. As the suppressor grid becomes more negative, the current partition ratio, $k_{cp} = I_p/I_{g2}$, decreases and a still smaller proportion of electrons reach the plate. As a consequence, the slope of the characteristics decreases.

The comparison of the characteristics given in Fig. 4-13*a* and *b* shows that despite the difference in the physical processes on which

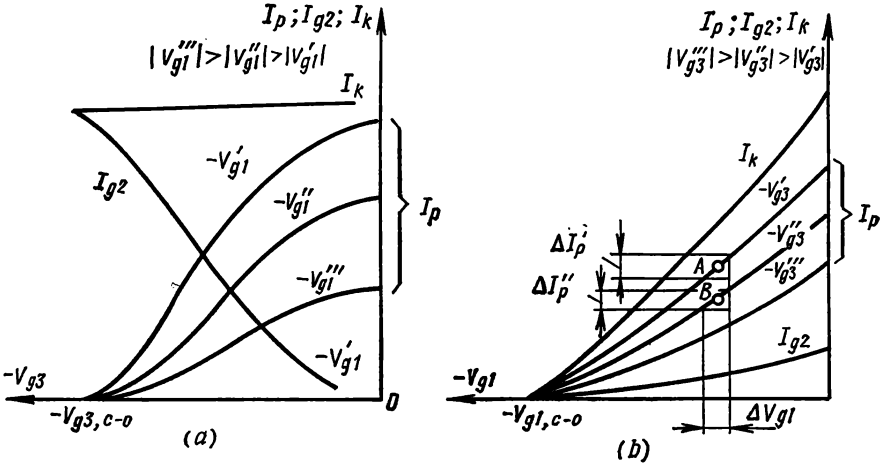


Fig. 4-13. Characteristics illustrating double-grid current control in the pentode

these relations are based, plate current in the pentode may be controlled by varying the voltages at the control and the suppressor grids. If, for example, the control grid is used to control plate current, this process may be regulated by varying the slope of the $I_p = \varphi_1 (V_{g1})$ characteristic, for which purpose we vary the suppressor-grid voltage.

4-3. PARAMETERS OF MULTIELECTRODE TUBES

In addition to the no-load parameters of the plate circuit, multielectrode tubes are also described in terms of parameters telling how effective V_{g1}, V_{g2} and V_p are in controlling I_{g2} (that is, the no-load parameters of the screen-grid circuit). In the pentode, the parameters associated with double-grid current control are also of interest.

No-Load Parameters of the Plate Circuit. As with the triode, in multielectrode tubes these parameters tell how much the electrode voltages affect the plate current.

The transconductance relates the tetrode or pentode plate current and control-grid voltage, V_{g1}, with the other voltages kept constant

$$g_m = \frac{dI_p}{dV_{g1}}\Bigg|_{V_{g2}=\text{const and } V_p=\text{const (for the pentode, } V_{g3}=\text{const)}} \qquad (4\text{-}18)$$

As with the triode, the transconductance can geometrically be defined as the slope of the plate-grid characteristic, $I_p = f_1(V_{g1})$. For beam power tetrodes and pentodes, the transconductance ranges from 2 to 20 mA/V.

The dynamic plate resistance has the same meaning and can be found geometrically in much the same manner as in the triode. It is determined with control-grid and screen-grid voltages held constant

$$R_p = \frac{dV_p}{dI_p}\Big|_{V_{g1}=\text{const and } V_{g2}=\text{const (for the pentode, } V_{g3}=\text{const)}} \tag{4-19}$$

The dynamic plate resistance for pentodes is much higher than it is for triodes, and ranges from 200 kilohms to several megohms.

As in the triode, the no-load amplification factor tells how many times the control-grid voltage, V_{g1}, is more effective than the plate voltage, V_p, in controlling the plate current

$$\mu = -\frac{dV_p}{dV_{g1}}\Big|_{I_p=\text{const and } V_{g2}=\text{const (for the pentode, } V_{g3}=\text{const)}} \tag{4-20}$$

Since V_p has a smaller effect on plate current than it does in the triode, the amplification factor is considerably higher. For tetrodes, it ranges from about 100 to 700, and for pentodes it reaches several thousand.

In addition to the three main no-load parameters, use is sometimes made of the screen-grid no-load amplification factor, μ_{pg2}, which relates V_p and V_{g2} in controlling I_p

$$\mu_{pg2} = -\frac{dV_p}{dV_{g2}}\Big|_{I_p=\text{const and } V_{g1}=\text{const (for the pentode, } V_{g3}=\text{const)}} \tag{4-21}$$

In the tetrode, V_{g2} is $1/D_1$ as effective as V_{g1} in controlling plate current due to the screening action of the control grid. So, μ_{pg2} is always smaller than μ and ranges from 10 to 200.

It should be noted that plate current changes not so much because variations in V_p and V_{g2} affect the space charge near the cathode as due to changes in the current partition ratio, k_{cp}.

No-Load Parameters of the Screen-Grid Circuit. Sometimes, a load may be connected in the screen-grid circuit. Then, it is important to know the screen-grid no-load parameters which relate I_{g2} to the tube voltages.

The screen-grid transconductance

$$g_{m\text{-}g2} = \frac{dI_{g2}}{dV_{g1}}\Big|_{V_{g2}=\text{const and } V_p=\text{const (for the pentode, } V_{g3}=\text{const)}} \tag{4-22}$$

relates the screen-grid current to V_{g1}. Graphically, it is given by the slope of the $I_{g2} = f_3(V_{g1})$ characteristic. The screen-grid transconductance, g_{m-g2}, is usually smaller than g_m because, under conditions of electron interception, I_{g2} and its absolute increment

caused by a change in cathode current are lower than I_p and the respective increment.

The dynamic screen-grid resistance

$$R_{g2}=\frac{dV_{g2}}{dI_{g2}}\Bigg|_{V_{g1}=\text{const and } V_p=\text{const (for the pentode, } V_{g3}=\text{const)}} \quad (4\text{-}23)$$

relates V_{g2} and the screen-grid current. Graphically, it can be found as the cotangent of the angle made by the $I_{g2} = f_4 (V_{g2})$ characteristic and the x-axis. R_{g2} is approximately the same as R_p in the triode.

The screen-grid no-load amplification factor

$$\mu_{g2g1}=-\frac{dV_{g2}}{dV_{g1}}\Bigg|_{I_{g2}=\text{const and } V_p=\text{const (for the pentode, } V_{g3}=\text{const)}} \quad (4\text{-}24)$$

relates V_{g2} and V_{g1} in controlling the screen-grid current. The screen-grid amplification factor varies within the same limits as the no-load amplification factor of the triode.

Parameters for Double-Grid Current Control in the Pentode. These include the suppressor-grid transconductance, the transconductance and the double-grid current control factor.

The *suppressor-grid transconductance* is the slope of the $I_p = \varphi (V_{g3})$ characteristic

$$g_{m3}=\frac{dI_p}{dV_{g3}}\Bigg|_{V_p,\ V_{g1} \text{ and } V_{g2} \text{ remaining constant}} \quad (4\text{-}25)$$

Depending on V_p, V_{g1} and V_{g2}, the suppressor-grid transconductance ranges from 0.1 to 6 mA/V.

The *transconductance* g_{m1} gives the slope of the $I_p = \varphi_1 (V_{g1})$ characteristic

$$g_{m1}=\frac{dI_p}{dV_{g1}}\Bigg|_{V_p,\ V_{g2} \text{ and } V_{g3} \text{ remaining constant}} \quad (4\text{-}26)$$

In contrast to the transconductance of the plate circuit (g_m), g_{m1} tells how effective V_{g1} is in controlling the plate current at $V_{g3} \neq 0$ (as a rule, $V_{g3} < 0$).

The *double-grid current control factor*

$$K_d=\partial^2 I_p/\partial V_{g1}\partial V_{g3}=\partial g_{m1}/\partial V_{g3}=\partial g_{m3}/\partial V_{g1} \quad (4\text{-}27)$$

inducates how much the control-grid voltage affects the suppressor-grid transconductance.

The Effect of Current Partition on the No-Load Parameters. In multielectrode tubes, we cannot consider the effect of the tube voltages on I_p and I_{g2} in terms of only their effect on the space charge near the cathode and, as a consequence, on the total stream of electrons moving from the cathode. We should also take into account variations in the current partition ratio.

The transconductance, g_m, is found by determining the increment in plate current, ΔI_p, caused by an increment in control-grid voltage, ΔV_{g1}. This, however, affects not only plate current, but also screen-grid current. So the increment in cathode current due to the change in V_{c1} as caused by V_{g1} is the sum of ΔI_p and ΔI_{g2}. The effect of V_{g1} on cathode current is expressed in terms of the cathode transconductance

$$g_{m\text{-}k} = \left.\frac{dI_k}{dV_{g1}}\right|_{V_p \text{ and } V_{g2} \text{ remaining constant}} \tag{4-28}$$

The cathode transconductance, g_{m-k}, may be related to g_m and g_{m-g2} by using Eqs. (4-13) and (4-14)

$$g_{m\text{-}k} = \frac{\partial}{\partial V_{g1}}\left(\frac{k_{cp}+1}{k_{cp}} I_p\right) = \frac{k_{cp}+1}{k_{cp}} \times \frac{\partial I_p}{\partial V_{g1}} = \frac{k_{cp}+1}{k_{cp}} g_m \tag{4-29}$$

$$g_{m\text{-}k} = \frac{\partial}{\partial V_{g1}}[(k_{cp}+1) I_{g2}] = (k_{cp}+1)(\partial I_{g2}/\partial V_{g1}) = (k_{cp}+1) g_{m\text{-}g2} \tag{4-30}$$

This suggests that g_{m-k} is always higher than g_m and g_{m-g2}. Referring to Eq. (4-29), other things being equal, g_m in the tetrode and pentode is always somewhat smaller than that in the triode $\left(\frac{k_{cp}+1}{k_{cp}} > 1\right)$ because, in contrast to the triode, in multielectrode tubes, $\Delta I_p < \Delta I_k$ (at $V_{g1} < 0$). This difference increases with decreasing current partition ratio.

As to the dynamic plate resistance of multielectrode tubes, we should keep in mind that variations in plate current accompanying variations in V_p are brought about by changes in the current partition ratio rather than by changes in I_k. The screening effect of the grids in the tetrode, and especially in the pentode, is so great that variations in V_p affect cathode current very little. At the same time, variations in plate current may be considerable because the electrons previously intercepted by the grid join the stream of electrons travelling to the plate.

As important is the effect of current partition on the no-load amplification factor, μ, which relates V_p and V_{g1} in controlling plate current irrespective of variations in I_k. The no-load amplification factor takes account of current partition in a tube. This can be demonstrated as follows. When determining the amplification factor, we note ΔI_p caused by ΔV_{g1}, then compensate for this increment by adjusting the plate voltage. The increment ΔI_p brought about by an increase in V_{g1} is attributed solely to a change in V_{c1} and, as a consequence, to a rise in the cathode current. The increment

ΔI_p is compensated for by V_p, especially in the pentode, mainly owing to variations in k_{cp}.

Now, let us recall the physical meaning of the tube penetration factor, D. As in the triode, in multielectrode tubes the penetration tells how far the lines of force of the plate field extend to the cathode and how much this field affects the space charge near the cathode and, as a consequence, cathode current as compared with the control-grid field

$$D = D_1 D_2 D_3 = -dV_{g1}/dV_p \quad I_k = \text{const}$$

When we discussed the relations between μ and D in the triode, we learned that $\mu = 1/D$ only if $I_p = I_k$ ($V_g < 0$ and $I_g = 0$). In multielectrode tubes, I_p is usually other than I_k, so $\mu \neq 1/D$.

This is well illustrated by Fig. 4-14*a*. Let us suppose that, when determining μ and D, we change V_{g1} by the same amount, ΔV_{g1}. Then, $\Delta I_k > \Delta I_p$. It takes a larger change in the plate voltage to compensate for ΔI_k than ΔI_p. In the latter case, the plate current varies not only because of the cathode current, but also because of variations in the current partition ratio. Thus

$$\Delta V_{pD}/\Delta V_{g1} > \Delta V_{p\mu}/\Delta V_{g1} \tag{4-31}$$

and, as a consequence,

$$1/D > \mu \tag{4-32}$$

In view of this, Eq. (3-42) which gives the tube equation of the triode does not hold for multielectrode tubes. For the tetrode and pentode, only Eq. (3-41) still holds

$$g_m R_p = \mu$$

Determining No-Load Parameters from Tube Characteristics. As already noted, data sheets for multielectrode tubes specify only families of plate characteristics. The characteristic triangles used when determining the no-load parameters of the triode (see Sec. 3-3) can hardly be applied to the plate characteristics of multielectrode tubes.

In the characteristics given by data sheets, adjacent curves run not so close to one another as it is specially shown in Fig. 4-14*a*, and the straight line $I_p = \text{const}$ drawn from the specified Q-point does not generally cross the adjacent characteristic within the operating region (operation under conditions of electron interception). So, the dynamic plate resistance of a tube, $R_p = \Delta V_p/\Delta I_p$, is determined by specifying an increment in plate voltage, ΔV_p, and finding the respective increment in plate current, ΔI_p (Fig. 4-14*b*). The transconductance can easily be found by drawing the straight line $V_p = \text{const}$ and determining $\Delta I'_p/g_m = \Delta I'_p/\Delta V_{g1}$, where $\Delta V_{g1} = |V''_{g1}| - |V'_{g1}|$. The no-load amplification factor is given by Eq. (3-41): $\mu = g_m R_p$.

The no-load parameters for double-grid current control can be determined as shown in Fig. 4-13. On finding g_{m1} at points A and B, we obtain $K_d = (g_{m1A} - g_{m1B})/|V'''_{g3} - V''_{g3}|$.

Interelectrode Capacitances. As already noted, each pair of electrodes in a tube may be regarded as the plates of a capacitor forming an interelectrode capacitance. Accordingly, the pentode has ten interelectrode capacitances. These capacitances, however, differ in their effect on the performance of a tube. An important capacitance is that between the plate and the control grid, C_{pg1}, that is, between the input and output circuits of the pentode. In multielectrode

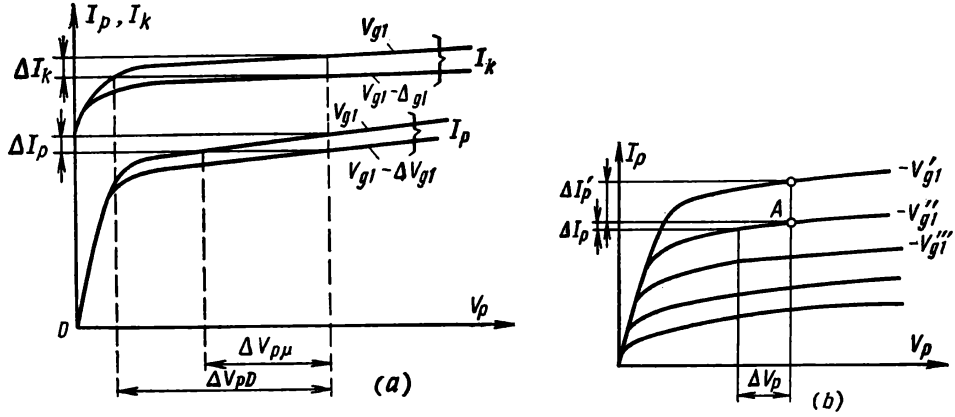


Fig. 4-14. Determining no-load parameters from characteristics
(*a*) determining μ and D; (*b*) determining g_m and R_p

tubes, the screen and suppressor grids reduce this capacitance to hundredths or thousandths of a picofarad.

Also important are the control grid-to-cathode and the plate-to-cathode capacitances shunting the input and output circuits, respectively. Accordingly, these capacitances representing the sums of several interelectrode capacitances are called the input and output capacitances of the pentode.

The input capacitance is formed by C_{g1k}, C_{g1g2} and C_{g1g3}, that is, by the capacitances between the control grid and the tube electrodes grounded for the a.c. component: the cathode, the screen grid and the suppressor grid connected to the cathode. C_{g1g3} is usually much smaller than the other two capacitances because the control grid is screened from the suppressor grid by the sufficiently dense screen grid. So, we may write for the input capacitance

$$C_{in} = (\text{approx.})\ C_{g1k} + C_{g1g2} \tag{4-33}$$

The output capacitance of the pentode is

$$C_{out} = (\text{approx.})\ C_{pk} + C_{pg3} + C_{pg2} \tag{4-34}$$

As a rule, the interelectrode capacitances are determined as sums of the capacitances between the electrodes proper and the capacitances between the electrode leads.

The input and output capacitances markedly affect the performance of a tube at radio frequencies because, as the frequency rises, the reactances of these capacitances fall. As a result, an increase in the signal frequency brings about a decrease in the input and output resistances of the tube, shunted by C_{in} and C_{out}.

4-4. CONSTRUCTION AND TYPES OF TETRODES AND PENTODES

Pentodes are the most commonly used tubes. They are employed in radio-frequency amplifiers, audio-frequency power amplifiers, oscillators, etc.

Beam power tetrodes are finding more limited application. They are mainly used in audio-frequency amplifiers, power amplifiers and oscillators.

The performance of a pentode operating in a radio-frequency amplifier is mostly affected by the transfer capacitance, C_{pg1}, which is the sum of the capacitance between the electrodes proper (the control grid and the plate) and the capacitance between their leads. So, in r.f. pentodes one seeks to reduce not only the capacitance between the plate and the control grid, for which purpose the screen grid wires are spaced closer apart, but also the capacitance between the leads of these electrodes by using shields inside and outside the tube. A.F. pentodes and beam tetrodes mainly used as power amplifiers in driver and output amplifier stages differ in construction and parameters from r.f. pentodes. For audio amplifiers, the value of C_{pg1} is of minor importance. In beam tetrodes and a.f. pentodes, C_{pg1} ranges from about 0.5 to 1 pF. Here, more coarse grids may be used to reduce I_{g2}.

Tetrodes and pentodes operating in output amplifier stages handle high currents because the purpose of output amplifier stages is to boost the signal in power. So, output stages use tubes with high-power cathodes whose emission may appreciably heat the grids. This may give rise to thermionic emission from the control grid. To avoid this, the control grid is sometimes plated with a thin layer of gold, and heat sinks are welded to its supports to improve heat abstraction. In high-power output tubes, some electrons accidentally hitting the envelope may cause secondary emission from its walls. So, the internal surface of such tubes is coated with graphite, and the edges of the plate are screened by a shield.

The construction of some multielectrode tubes is shown in Fig. 4-15.

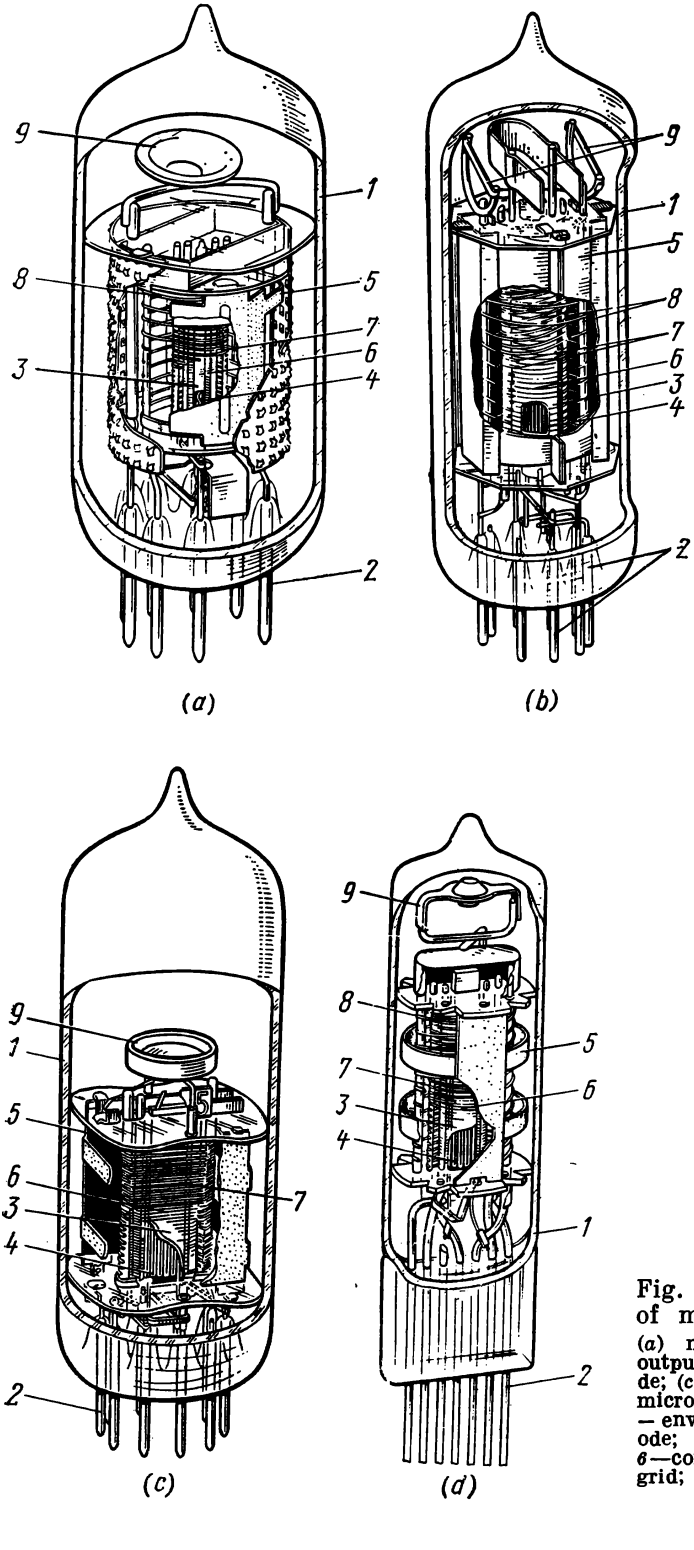


Fig. 4-15. Construction of multielectrode tubes (*a*) miniature pentode; (*b*) output-stage miniature pentode; (*c*) miniature tetrode; (*d*) microminiature pentode; *1*—envelope; *2*—pins; *3*—cathode; *4*—heater; *5*—plate; *6*—control grid; *7*—screen grid; *8*—suppressor grid; *9*—getter

4-5. VACUUM TUBES FOR WIDEBAND AMPLIFIERS

This group of tubes includes some tetrodes and pentodes intended to amplify signals with a wide frequency spectrum, such as short-duration pulse signals, and also tubes with special electrode structures.

To amplify a broadband signal without distorting its waveform, a tube amplifier stage should give the same amplification to the whole range of frequencies. In a tube amplifier, however, the plate load resistor, R_L, is shunted by the output capacitance of the tube and also by the input capacitance of the succeeding amplifier stage. So, the load impedance is complex and frequency-dependent

$$Z_L(\omega) = R_L/[1 + i\omega(C_{out} + C_{in})R_L] \qquad (4\text{-}35)$$

By Eq. (3-70), the stage gain in terms of voltage is

$$k_v = \mu/(1 + R_p/R_L) = g_m R_p R_L/(R_L + R_p)$$

In the pentode, as a rule, $R_p \gg R_L$, so we may write

$$k_v = (\text{approx.})\, g_m R_L \qquad (4\text{-}36)$$

or, substituting Eq. (4-35) into this expression,

$$k_v(\omega) = (\text{approx.})\, g_m R_L/[1 + i\omega(C_{out} + C_{in})R_L] \qquad (4\text{-}37)$$

This indicates that $k_v(\omega)$ decreases with increasing frequency. Thus, to raise the voltage gain at radio frequencies without distorting wideband signals, we should increase the transconductance and reduce the input and output capacitances. To meet the latter requirement, wideband tetrodes and pentodes are built into envelopes that have no bases.

The transconductance may be raised to as high as 30 mA/V by using a control grid with closely spaced wires and placing it closer to the cathode. To avoid the danger of short-circuits between the grid wires and the cathode due to the heating and sagging of the wires, use is made of a frame grid obtained by winding wire about five micrometres in diameter with a pitch of about 60 μm on a rectangular frame. With this arrangement, however, the grid-cathode spacing cannot be made less than 30 or 40 μm.

Screened-Cathode Tubes. Another way to increase the transconductance and reduce the input capacitances is to use a *cathode screen* (or *shield*) interposed between the cathode and the control grid and held at a low positive potential (several volts). The control grid is maintained at a low negative potential of several volts. The potential distribution in such a tube is shown in Fig. 4-16. In this tube, there is a zero-potential region or a virtual cathode at a very small distance, Δx, from the control grid. Owing to the small Δx, the space charge in the virtual cathode region can be controlled rather effectively, and the transconductance may reach 50 mA/V. As is

seen, the screened-cathode tube operates as if the distance between the control grid and the cathode were reduced. The cathode screen is placed sufficiently far from the cathode, so the input capacitance of the tube is one half or even one third that of the ordinary pentode. An increase in transconductance and a decrease in input capacitance bring about an increase in amplification factor.

There is a third, screen, grid. As in the beam power tetrode, the dynatron effect is eliminated by using special shields held at zero potential with respect to the cathode.

The static characteristics of the Soviet-made 6Ж22П pentode using a cathode screen are shown in Fig. 4-17.

Secondary-Emission Tubes. A high transconductance is secured in secondary-emission tubes, or dynatrons. A dynatron has secondary-emission cathodes, called *dynodes*. The construction of a dynatron is shown in Fig. 4-18.

The tetrode section of the tube is of usual construction. The electrons which have passed through the screen grid are deflected by beam-forming electrodes, make up current I_{d1}, and strike a dynode. Dynodes are usually made of a material with a secondary emission ratio, σ, ranging from 4 to 5 (for example, nickel activated by cesium oxide or a mixture of alkali-earth metal oxides) and meet the general requirements for plates. The dynode is held at a lower potential than the plate, so all secondary electrons make for the plate and constitute current I_{d2}. If all primary electrons reach the dynode, the plate current is

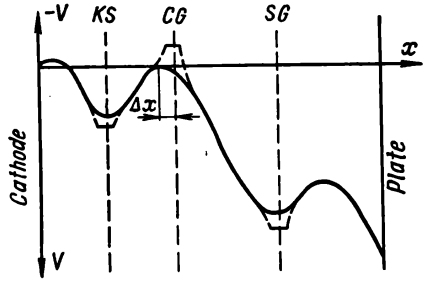


Fig. 4-16. Potential distribution in a screened-cathode tube

$$I_{pd} = I_{d2} = \sigma I_{d1} = \sigma I_p \qquad (4\text{-}38)$$

where I_p is the plate current in the ordinary pentode.

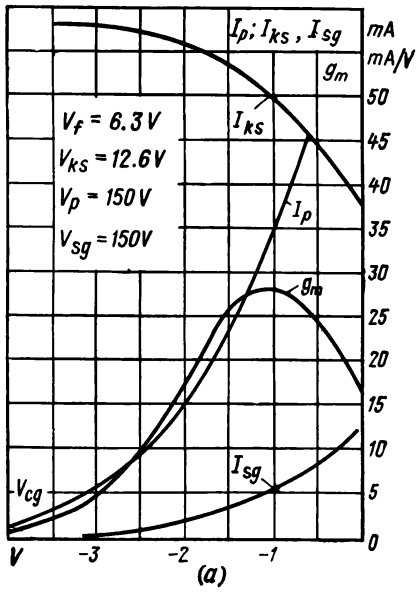
In practice, however, some of the primary electrons do reach the plate and constitute current I_{pd1}. In this case

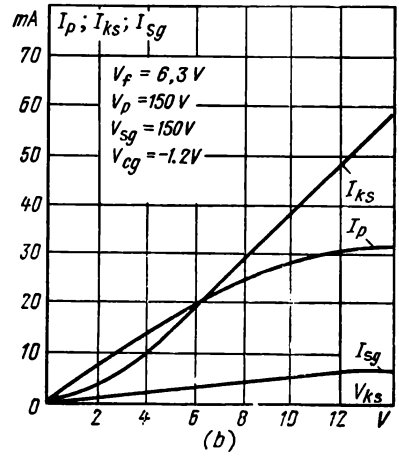
$$I_{pd} = I_{pd1} + \sigma I_{d1} \qquad (4\text{-}39)$$

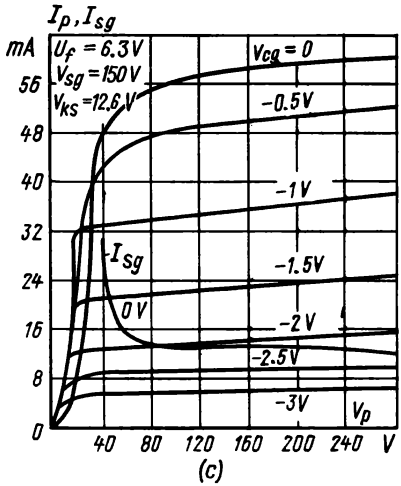
The characteristics of a Soviet-made type 6B1П dynatron (Fig. 4-19) demonstrate some characteristic features of this tube. Referring to Fig. 4-19*a*, the plate current, I_{pd}, and the dynode current $I_d = I_{d2} - I_{d1}$, oppose each other (provided that $\sigma > 1$). So in an amplifier, the voltages across the resistors connected in the circuits of these electrodes may be 180° out of phase.

In this tube, the transconductances of the plate and dynode circuits are high, but $g_{m\text{-}d} = dI_d/dV_{g1}$ is negative.

As can be seen from Fig. 4-19*b*, I_{pd} and I_d reach their peaks at a certain dynode voltage, V'_d (as a rule, $V'_d \approx 0.6V_p$). At lower values of V_d, I_{pd} and I_d decrease because some electrons return to the screen grid, and also because the secondary-emission ratio falls.







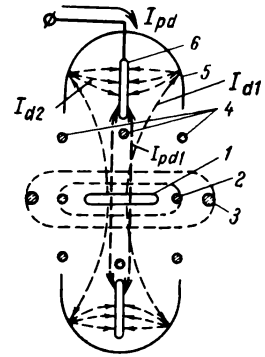


Fig. 4-17. Characteristics of a screened-cathode tube

(*a*) plate-grid characteristics; (*b*) currents versus cathode-screen voltage; (*c*) plate characteristics

Fig. 4-18. Construction of a secondary-emission tube

1—cathode; *2*—control grid; *3*—screen grid; *4* — beam-forming screen; *5*—dynode; *6*—plate

When V_d is low ($\sigma < 1$), the plate current is mainly made up of primary electrons: I_{pd} = (approx.) I_{pd1}. As V_d rises, some secondary electrons return to the dynode and I_{pd} decreases. The dynode current, I_d, also falls because the space charge forces the secondary electrons to fall back on the dynode.

The plate characteristics of the tube also have a peculiar shape (Fig. 4-19*c*). At low V_p, $I_{pd} = 0$ because all electrons are intercepted by the dynode. The plate voltage does not affect the dynode

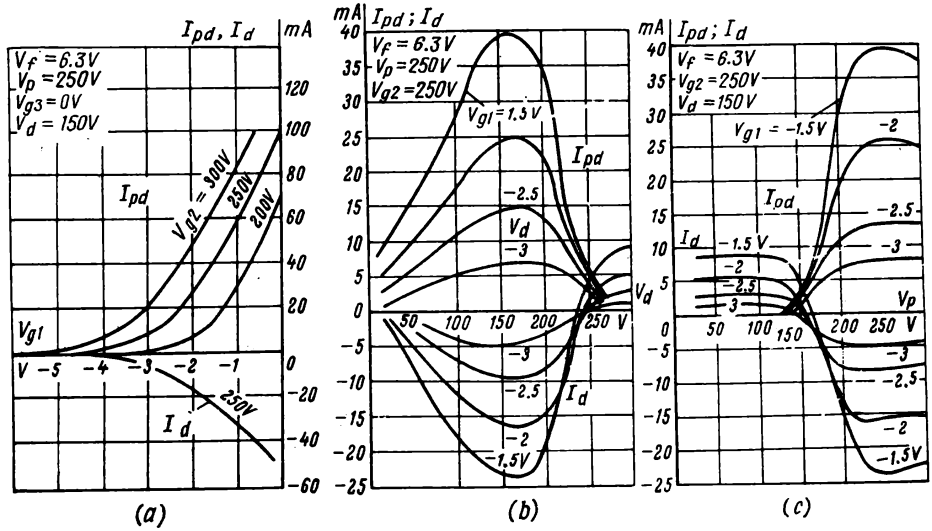


Fig. 4-19. Characteristics of a secondary-emission tube
(*a*) plate-grid characteristics; (*b*) currents versus dynode voltage; (*c*) plate characteristics

current. As V_p rises, I_{pd} increases at first due to the interception of some primary electrons, and then owing to secondary electrons. As this takes place, I_d falls and reverses.

Owing to secondary emission, the transconductance of these tubes is much higher than that of ordinary pentodes. For example, the transconductance of the type 6B1П dynatron is 28 mA/V; those of Soviet-made type 6B2П and 6B3П pulse tubes reach 300 mA/V in pulse work ($\tau = 1$ μs).

4-6. RADIO-FREQUENCY TUBES

At frequencies above 100 MHz, the performance of vacuum tubes is affected by some factors which have not been taken into account before.

Transit-Time Effects. At low frequencies, the transit time (the time of flight of an electron between a pair of electrodes), τ, is

considerably shorter than the period, T, of the alternating voltage applied to the tube. As a consequence, the phase of the voltage can be considered constant during the time τ. At radio frequencies, however, τ is an appreciable part of T. Accordingly, during the transit time plate voltage might markedly change in phase, so an additional phase shift might be introduced between the plate current and plate voltage.

It is usual to relate the transit time to the period of the applied alternating voltage in terms of what is known as the *transit angle*, θ

$$\theta = 2\pi\tau/T = 2\pi f\tau = \omega\tau$$

In a diode, the transit time depends on the plate voltage and the distance between the cathode and the plate. In a diode with a plane cathode and a plane-parallel plate, where an electron starts out at zero initial velocity, its transit time is defined by an equation given in reference [1]

$$\tau = (\text{approx.})\ 5 \times 10^{-8} r_p V_p^{-1/2} \qquad (4\text{-}40)$$

As an example, let us find the transit time and transit angle for a planar diode with the cathode and the plate spaced the distance $r_p = 1$ cm apart and at a plate voltage of 100 V. Substituting these values into Eq. (4-40), we get τ = (approx.) 5.1×10^{-9} s. If the frequency of the applied alternating voltage is 100 MHz, the transit angle is $\theta = 5.1 \times 10^{-9}\ 2\pi \times 10^8$ = (approx.) π. This means that, if the electron starts out from the cathode when the alternating voltage is at its positive peak, it will reach the plate when its phase has changed by 180°, that is, when the alternating voltage is at its negative peak.

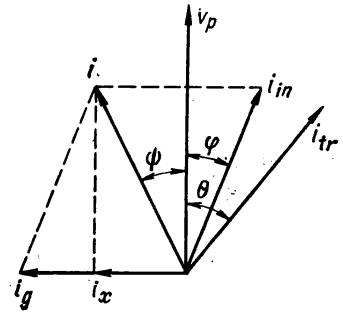


Fig. 4-20. Vector diagram of voltages and currents in a diode

The vector diagram of voltages and currents in a diode with a radio-frequency alternating voltage applied between its plate and cathode is shown in Fig. 4-20. The current made up by the moving electrons (transfer current, i_{tr}) lags behind the alternating plate current due to the electron time lag. In the diagram, this phase shift is shown by the angle θ between the plate voltage vector, v_p, and the transfer current vector, i_{tr}, near the plate. The instantaneous value of the induced current is equal to the average transfer current at the given time. So, the phase shift angle φ between the induced current (i_{in}) and v_p is somewhat smaller than θ. The capacitive current, i_C, flowing through the plate-to-cathode capacitance leads v_p by the angle $\pi/2$. The total current, i, in the external circuit is equal

to the sum of the capacitive current and the induced current. Thus, the phase shift between the current in the external circuit of the diode and the applied voltage is ψ and, as a consequence, the impedance of the diode at radio frequencies is a complex quantity

$$Z_\tau = R_\tau + iX_\tau$$

Any tube containing grids may be regarded as a tube consisting of a number of vacuum gaps each made up by a pair of adjacent electrodes. In the pentode, for example, we may distinguish the

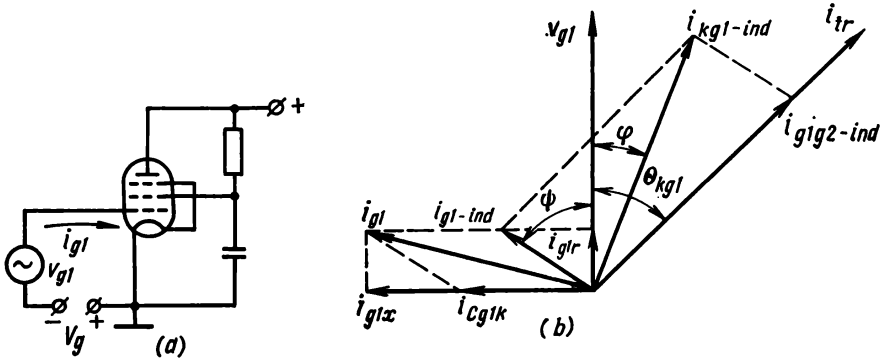


Fig. 4-21. (*a*) Input circuit of the pentode and (*b*) vector diagram of currents and voltages

gaps between the cathode and the control grid, between the control grid and the screen grid, and so on. Any of these sections obeys the laws derived for the diode.

The electron time lag mostly affects the space between the cathode and the control grid. In triodes and pentodes, the control grid is usually held at a negative potential, and the electrons emitted by the cathode move at a low velocity. On passing through the control grid, the electrons find themselves in the accelerating field of the screen grid, so they move to the plate at a considerable velocity. As is seen, the travel from the cathode to the control grid usually takes the largest part of the transit time.

Let us discuss the vector diagram of voltages and currents in the input circuit of the tube (Fig. 4-21). In the cathode-control grid space near the control grid, the transfer current, i_{tr}, lags behind the control grid voltage, v_{g1}, by the angle θ_{kg1} due to the electron time lag. In the cathode control grid space, the induced current $i_{kg1\text{-}in}$, equal to the average transfer current, lags behind the voltage by the angle $\varphi < \theta_{kg1}$. In the space between the control grid and the screen grid, the electrons move at a high velocity, their transit angle is small, and the phase shift between i_{tr} and v_{g1} during the time τ_{g1g2} may safely be deemed unchanged. So, $i_{g1g2\text{-}in}$ induced

by the electrons in the first-grid lead when they travel to the second grid is in phase with i_{tr}. However, $i_{kg1\text{-}in}$ and $i_{g1g2\text{-}in}$ are in antiphase because the former is due to the electrons approaching the grid, and the latter, to the electrons moving away from the grid. The resultant current is thus equal to the difference between these currents

$$i_{g1\text{-}in} = i_{kg1\text{-}in} - i_{g1g2\text{-}in} \tag{4-41}$$

In addition to the induced current, the input circuit carries a current flowing through the cathode-to-control grid capacitance, C_{g1k}. Accordingly, the total current in the external circuit, i_{g1}, is equal to the sum of the induced current, $i_{g1\text{-}in}$, and the capacitive current, i_{Ckg1}, leading v_{g1} by $\pi/2$.

Referring to the diagram, i_{g1} leads v_{g1} by an angle smaller than $\pi/2$, which means that the input admittance of the tube is a complex quantity. The presence of the conductance current shows that the input admittance contains a real component, that is, some of the signal power is dissipated by the input circuit of the tube. It can be shown that in a pentode the real component of the input admittance (the conductance) depends on the transit angle. According to an equation given in reference [1]

$$G_{in\text{-}\tau} = (1/20)\, g_{m\text{-}k}\omega^2\tau_{kg1}^2 \tag{4-42}$$

where $g_{m\text{-}k} = dI_k/dV_{g1}$.

As can be seen from Fig. 4-21, the vector of the reactive component of the input current, i_{g1x}, exceeds that of the capacitive current, i_{Ckg1}. From a comparison with the vector diagram for the diode (Fig. 4-20) it is seen that in the diode the reactive component, i_x, of the input current, i, is smaller than the capacitive current, i_C. This difference is due to different phase shifts between the induced current and voltage. In the diode, i_{in} lags behind the applied r.f. voltage, which is equivalent to placing in circuit a certain inductive reactance which reduces the capacitive current in the tube. In multielectrode tubes, the currents $i_{kg1\text{-}in}$ and $i_{g1g2\text{-}in}$ induced in the grid circuit oppose each other and lead the applied voltage, which is equivalent to an increase in the input capacitance. Accordingly, the equivalent capacitance of the diode is only a part of C, and the equivalent input capacitance of multielectrode tubes is greater than C_{kg1}.

The Effect of Distributed Reactances. In multigrid tubes, the number of distributed reactances is rather large (Fig. 4-22). However, not all elements shown in Fig. 4-22 affect the tube performance equally. The effect of the interelectrode capacitances on the tube performance at radio frequencies has been discussed in Sec. 4-4. The tube performance is also affected to the same degree by the distributed inductances of the leads connecting the electrodes to the

pins. In conventional receiving tubes, these inductances are rather small, being a fraction of a microhenry. However, at frequencies of about several thousand megahertz, the inductive reactances reach several tens of ohms and become comparable with the reactances of the interelectrode capacitances or, in the diode, with the dynamic plate resistance. So, at radio frequencies, the lead inductances markedly affect the characteristic admittances of the tube. This effect is felt in cold tubes where the inductances force the reactive components of the characteristic admittances to change, as well as in hot

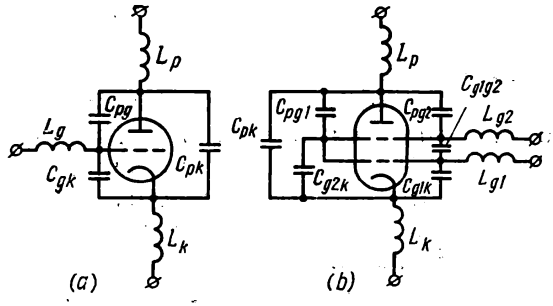


Fig. 4-22. Distributed reactances

tubes where the phase shift between the current in and the voltage across an inductance may cause changes in the real components of the admittances.

For example, let us consider how much the inductance of the cathode lead affects the input admittance of the pentode. Suppose that an alternating voltage, v_{in}, made up of the voltage v_{g1k} across C_{g1k} and the voltage drop v_L across the inductance of the cathode lead, L_k, is applied between the control grid and ground. We construct a vector diagram (Fig. 4-23) and determine the input admittance

$$Y_{in} = i_{in}/v_{in} \qquad (4\text{-}43)$$

where i_{in} is the current in the input circuit of the tube.

The voltage across the control grid-to-cathode capacitance, v_{g1k}, gives rise to a cathode current, i_k, which is in phase with this voltage. The cathode current flowing in L_k establishes a voltage drop, $v_L = i\omega L_k i_k$, across it, which leads the cathode current by $\pi/2$.

The input voltage equal to the sum of voltages, $v_{in} = v_{g1k} + v_L$, is obtained by combining the respective vectors in the vector diagram.

The input current is equal to the sum of the currents flowing in C_{g1k} and C_{g1g2}: $i_{in} = i_{g1k} + i_{g1g2}$.

The voltage v_{g1k} is applied across the capacitance C_{g1k}, so, as a consequence, the current, i_{g1k}, flowing in this capacitance leads the grid-to-cathode voltage by $\pi/2$.

The voltage applied to C_{g1g2} is the input voltage, v_{in}, because the screen grid is grounded for radio frequencies, so i_{g1g2} leads v_{in} by $\pi/2$.

Adding together i_{g1k} and i_{g1g2}, we get the vector i_{in}. Referring to the diagram shown in Fig. 4-23, this current leads the input voltage by the angle $\psi < \pi/2$ and, as a consequence, the input admittance has a real (conductance) component.

The input admittance is defined by an equation given in reference [1]

$$Y_{in} = G_{inL} + iB_{inL} = \omega^2 C_{g1k} L_k g_{m-k} + i\omega (C_{g1k} + C_{g1g2}) \qquad (4\text{-}44)$$

As is seen, the real component of the input admittance depends on L_k and rises in proportion to the squared frequency of the applied voltage.

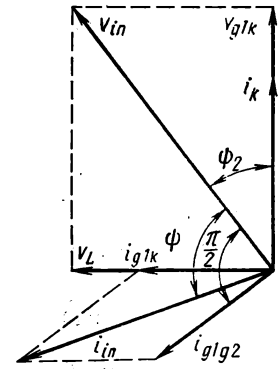


Fig. 4-23. Vector diagram illustrating the effect of the cathode-lead inductance

Because of L_k and other distributed reactances the input resistance of the pentode at 30 MHz may go down to about 10 kilohms, that is, approximately to one hundredth (at audio frequencies, $R_{in} \approx 1$ megohm).

The distributed reactances also affect the transconductance of the tube. Since the transit time is finite, the transfer current in the tube lags behind the voltage applied to the grid. Accordingly, the current in the plate circuit lags behind the input voltage by a certain angle θ_{kg1} (Fig. 4-21). This phase shift is aggravated by L_k because there is a phase shift, ψ_2, between the grid-cathode voltage and the input voltage (Fig. 4-23). Thus, at radio frequencies the transconductance is a complex quantity.

RF and Microwave Vacuum Tubes. At radio and microwave frequencies, the effect of transit time and distributed reactances can be eliminated or reduced by using tubes of special construction.

The simplest way to reduce the transit time is to reduce the spacing between the electrodes. This, however, increases interelectrode capacitances. So, in r.f. tubes, both the interelectrode spacing and the surface area of the electrodes are reduced. As a result, the transconductance is increased and the transit time is shortened, whereas the interelectrode capacitances remain almost the same.

As often as not, r.f. tubes have several parallel leads from the cathode, which considerably reduce the cathode inductance, L_k.

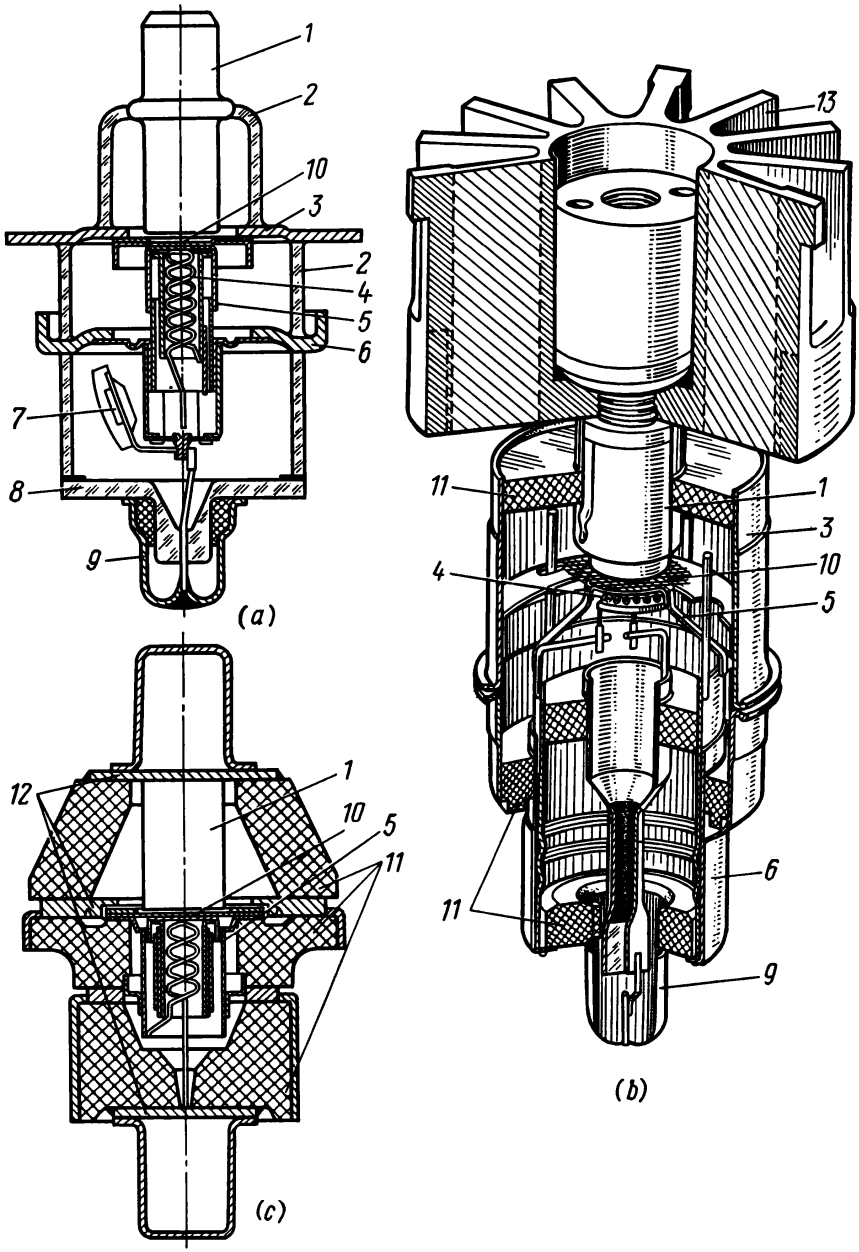


Fig. 4-24. Construction of microwave tubes

(*a*) lighthouse triode; (*b*) metal-ceramic triode; (*c*) titanium-ceramic triode; *1*—plate; *2*—glass insulator; *3*—grid lead; *4*—heater; *5*—cathode; *6*—cathode lead; *7*—getter; *8*—stem; *9*—heater lead; *10*—grid; *11*—ceramic parts; *12*—titanium washers; *13*—heat sink

At radio frequencies, use is sometimes made of triodes connected in a common-(grounded-) grid circuit. In this circuit, the control grid rather than the cathode is grounded. The inductance of the control-grid lead has the same effect on the input admittance of the tube as L_k in the grounded-cathode circuits. So, triodes intended for operation with the grounded grid contain several parallel control-grid leads or even solid disc seals.

Miniature and microminiature tubes show good performance at frequencies up to several hundreds of megahertz. At higher frequencies, their normal operation is upset due to an increase in the transfer admittance.

In the UHF and partly the SHF bands, use is made of *lighthouse* and *metal-ceramic tubes*. A lighthouse tube* (Fig. 4-24*a*) is an UHF vacuum diode or triode shaped like a lighthouse and having disc-sealed planar elements. The interelectrode spacings are very small, being about 50 μm. The plate is usually a massive rod. The plate, grid and cathode have disc leads; the leads are connected to coaxial lines which operate as resonant circuits in this frequency range. Thus, owing to small interelectrode spacings and also massive disc leads connected directly to resonant circuits, the effect of transit time and distributed reactances in lighthouse tubes is minimal.

Metal-ceramic tubes (Fig. 4-24*b*) also have planar electrodes whose leads are insulated not by glass envelopes as in lighthouse tubes, but by cylinders made of r.f. low-loss ceramic. Like lighthouse tubes, metal-ceramic tubes can conveniently be connected to coaxial resonant circuits owing to their disc-shaped leads.

The latest metal-ceramic tubes use titanium washers as coupling elements, which can easily be sealed to ceramic (Fig. 4-24*c*). These tubes have smaller interelectrode spacings (12 to 25 μm), lower lead inductances and smaller size than lighthouse tubes.

Lighthouse and metal-ceramic tubes can be used at frequencies of up to several thousands of megahertz.

4-7. SPECIAL-PURPOSE TUBES

Special-purpose tubes include vacuum tubes intended for operation in a narrow class of electronic devices (such as frequency converters, high-power oscillators, modulators, and so on), and having a certain characteristic feature (increased reliability, efficiency, etc.).

Multisection Tubes. These contain two or more groups of electrodes associated with independent electron streams within one envelope. Multisection (or multiple-unit) tubes include *double triodes*, *diode-triodes*, *triode-pentodes*, *triode-hexodes*, and so on.

---

* Also called disc-seal tube or megatron.— *Translator's note.*

Multisection tubes reduce the number of tubes, tube sockets, mounting elements and electronic components in apparatus. The manufacture of multisection tubes also has its advantages: the number of tube elements (envelopes, cathodes, heaters, etc.) and operations per tube is reduced.

In some multisection tubes, the various electrode groups are combined only physically; each section is functionally self-contained and has electrodes of its own, including the cathode. An illustrative example of such a tube is the Soviet-made type 6Ф3П triode-pentode whose triode section containing a separate cathode is often used as an oscillator while the pentode section serves as an output amplifier.

In some types of multisection tubes, the cathode is common to two or three electrode groups as, for example, in the Soviet-made type 6Г2 double diode-triode or the Soviet-made type 6И1П triode-heptode (Fig. 4-25*b*).

Frequency-Converter Tubes. These are intended for operation in superheterodyne receivers where the carrier frequency of the received r.f. signal is converted from its original value to a lower one, called the *intermediate frequency*, which is then amplified. Frequency conversion takes place in a mixer tube whose grids are fed with two voltages: the input signal and the local-oscillator signal so-called because it is supplied by an oscillator incorporated locally in the frequency converter. The local-oscillator frequency should be either higher or lower than that of the received signal. As a result, the electron stream of the mixer is affected by two voltages, one at frequency ω_s and the other at frequency ω_o, and the tube output yields a voltage whose frequency is equal to the difference of these two frequencies: $\omega_s - \omega_o = \omega_i$ or $\omega_o - \omega_s = \omega_i$.

Frequency conversion and generation of oscillations at ω_o, can be combined in a single tube known as the heptode, that is, a tube with five grids*, or in a dual vacuum tube (such as a triode-pentode and a triode-heptode). In the heptode, the cathode and the first two grids make up a triode used as the local oscillator; the remaining section serves as the mixer. The local-oscillator and mixer sections share a common electron stream.

In dual tubes, the triode section is the local oscillator, and the pentode, hexode or heptode section serves as the mixer. In these tubes, each section is associated with an independent electron stream.

Figure 4-25*a* shows the construction of the Soviet-made type 6А2П heptode. The triode section includes the cathode, the first (control) grid and the second grid which serves as the plate of this section.

Mixing is carried out by the pentode section whose cathode is the space charge near the third grid; the control grid is the third

* Hence, its other name is the pentagrid tube.— *Translator's note.*

(signal) grid which accepts the incoming signal, and the fourth grid acts as the screen grid. The fifth (suppressor) grid, as well as the screen grid, reduces the coupling between the plate and the signal grid circuits.

The Soviet-made type 6И1П triode-heptode shown in Fig. 4-25*b* is a more efficient converter tube. The triode located in the upper

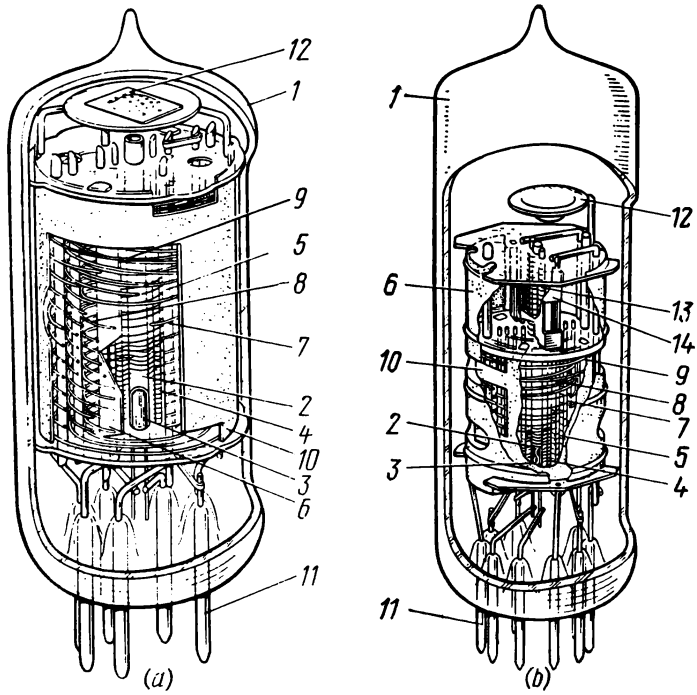


Fig. 4-25. (*a*) Heptode and (*b*) triode-heptode
1—envelope; *2*—cathode; *3*—heater; *4*—first heptode grid; *5*—second grid; *6*—screens; *7*—third grid; *8*—fourth grid; *9*—fifth grid; *10*—plate; *11*—pins; *12*—getter; *13*—triode grid; *14*—triode plate

part of the common cathode serves as the local oscillator. The local-oscillator voltage is applied to the third grid, and the signal voltage to the first grid of the mixer heptode. Owing to this construction, the influence of the mixer section on the local oscillator is completely eliminated. The third (signal) grid is separated from the plate by two grids.

The curves relating plate and screen-grid currents to control-grid and signal-grid voltages of the heptode are shown in Fig. 4-26. The curves $I_p = f_1(V_{g1})$ at $V_{g3} = \text{const}$ and $I_p = f_2(V_{g3})$ at $V_{g1} = \text{const}$ are similar to the respective curves of the pentode. The sharply changing slope of $I_p = f_2(V_{g3})$ is due to the construction of the third grid which, like that in the variable-μ tube, has variable

spacing between the wires. The third grid accepts a negative voltage whose value depends on the signal amplitude, and which automatically controls the amplification factor of the tube.

The shape of the curves $I_{g2+g4} = f_3(V_{g1})$ and $I_{g2+g4} = f_4(V_{g3})$ clearly shows the difference in the plate-current control effected by these two grids. As the first grid becomes less negative, both I_p and I_{g2+g4} increase. A fall in $-V_{g3}$ brings about a decrease in

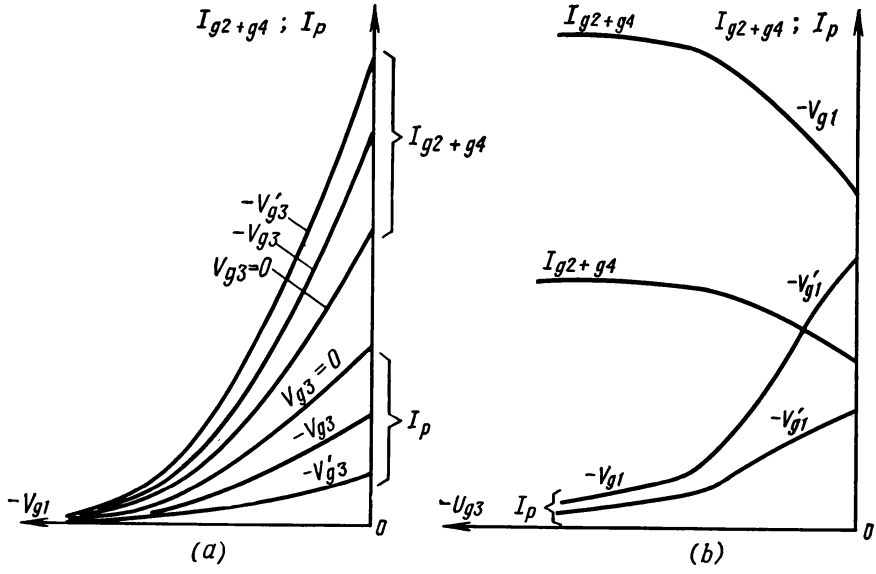


Fig. 4-26. Characteristics of the heptode (a) I_p and I_{g2+g4} as functions of first-grid voltage; (b) I_g and I_{g2+g4} as functions of third grid voltage

I_{g2+g4}, because a still smaller proportion of electrons return to the screen connected to the second grid.

Frequency-converter tubes are specified in terms of conventional no-load parameters whose values are within the limits for triodes and pentodes, respectively.

The performance of these tubes is expressed in terms of conversion parameters. The most important is the conversion transconductance, g_c, which is seen to be the rate of change of intermediate-frequency current amplitude with signal voltage amplitude

$$g_c = \frac{I_{m\text{-}i}}{V_{mg3}}\bigg|_{V_{g1}=\text{const}} \qquad (4\text{-}45)$$

or

$$g_c = K_d V_{mg1}/2 \qquad (4\text{-}46)$$

where

$$K_d = \partial^2 I_p / \partial V_{g1} \partial V_{g3} = \partial g_{m1} / \partial V_{g3} = \partial g_{m3} / \partial V_{g1}$$

is the double-grid current control factor defined by Eq. (4-27).

For some time past, frequency-converter tubes have been giving way to semiconductor devices.

High-Power Tubes. These include triodes, tetrodes, beam-power tetrodes and pentodes intended to generate CW or pulse signals with a power of a few tens of watts to several hundreds of kilowatts, and also tubes used in the modulators of high-power transmitters.

The physical principles of high-power tubes are similar to those of conventional triodes, tetrodes and pentodes. They differ in construction of electrodes, size and shape of envelopes.

The cathodes of such tubes ensure high emission currents up to hundreds of amperes. Transmitting tubes having a relatively low power (up to 1 kW) generally use oxide-coated cathodes. More powerful tubes have tungsten-carbide cathodes. The electrodes and leads in transmitting tubes handle voltages up to several tens of kilovolts. Because of this, the plate, the cathode and the grids are brought out to different sides of the envelope and sealed directly into glass. Their leads are usually made of multiconductor wire or substantial bars to reduce the resistance (Fig. 4-27*b*).

The plates of transmitting tubes are designed to dissipate high power. In low-power tubes, the plates are most often made of nickel; the plates of high-power tetrodes and pentodes are made of molybdenum, tantalum or graphite. For better heat abstraction, their surfaces are blackened or coated with titanium.

Plate cooling is very important in transmitting tubes. Natural plate cooling is utilized only in tubes with a power of not over 1 kW. High-power tubes use forced-air or liquid-cooled plates. In the former case, the massive plate lead carries a finned heat sink whose large surface area helps to dissipate heat (Fig. 4-27*a*). Sometimes, the sink is blown over by a jet of air from a fan. With water cooling, the plate is immersed in a running-water tank (Fig. 4-27*b*).

High-power transmitting tubes usually have a "right" plate-grid characteristic. Such tubes have a low cut-off potential, and the linear portion of the characteristic extends into the positive V_{g1} region.

High-power modulator tubes are similar to high-power transmitting tubes in construction, although their plate-grid characteristic is "left". Such a characteristic is necessary to amplify modulating signals without distortion.

Pulse modulating tubes use oxide-coated cathodes designed to handle high-density current pulses; these tubes should stand up well to high pulse voltage.

Efficient, Reliable and Long-Life Tubes. The efficiency of tubes is defined by the following ratio

$$E = g_m/P_o \tag{4-47}$$

where

$$P_o = I_f V_f + I_p V_p + I_{g2} V_{g2} \tag{4-48}$$

is the total power consumed by the tube, and g_m is the transconductance.

The efficiency of a tube can be raised by reducing filament current and voltage, plate voltage and screen-grid voltage, and by increa-

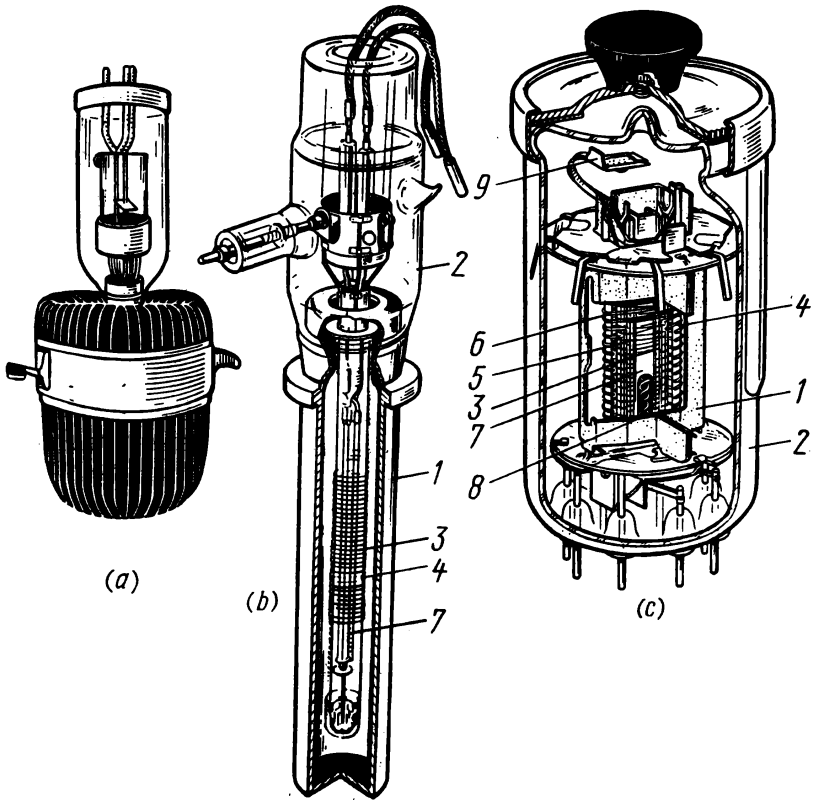


Fig. 4-27. High-power tubes
(*a*) transmitting triode with an air-cooled plate; (*b*) transmitting triode with a water-cooled plate; (*c*) transmitting pentode; *1*—plate; *2*—envelope; *3*—control grid; *4*—grid supports; *5*—screen grid; *6*—suppressor grid; *7*—cathode; *8*—heater; *9*—getter

sing, if possible, the transconductance. Efficient tubes include miniature and microminiature tubes having directly-heated cathodes and using a filament voltage of 1.2 V. These tubes use conventional wound grids.

High efficiency, shock and vibration resistance, and higher reliability are offered by tubes having rod electrodes. In these tubes, the electron stream is controlled by stiff metal rods placed instead of conventional grids and arranged in a certain manner between the filamentary cathode and the plate. A schematic section through such a tube is shown in Fig. 4-28*a*.

Cathode current is controlled by varying the potential at the control rods, G_1. When the rod electrodes are at zero potential, the

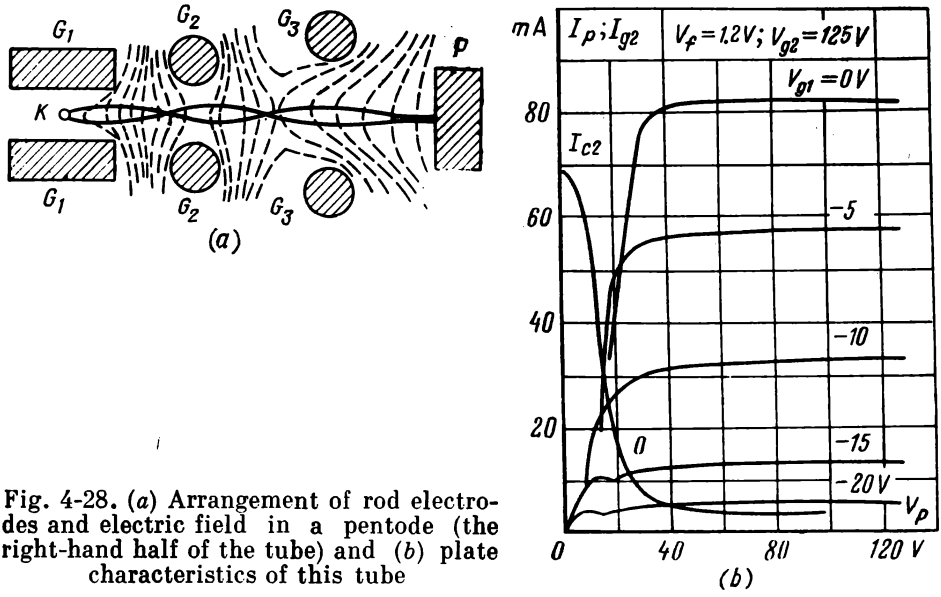


Fig. 4-28. (*a*) Arrangement of rod electrodes and electric field in a pentode (the right-hand half of the tube) and (*b*) plate characteristics of this tube

cloud made up by the electrons leaving the cathode surface is cylindrical in shape; when the rods are at a negative potential, it becomes elliptical. As is seen, plate current is controlled not only by changing the height of the potential barrier near the cathode, but also by varying the effective surface area of the cathode, that is, the region occupied by the space charge.

Between the grids, the configuration of the electric field helps to focus the stream of electrons travelling near the axis. Owing to the focused electron stream, the plate characteristics of tubes using rod electrodes have a very gradual operating region, and the transition from electron flyback to electron interception takes place at low plate voltages. Because of this, such tubes can be used at very low plate voltages. For example, the Soviet-made type 1Ж42A tube operates at $V_p = V_{g2} = 6V$.

Tubes using rod electrodes are usually of miniature and microminiature construction, with soft electrode leads. The plate is

brought out separately to the top of the envelope to reduce the interelectrode capacitances.

High-reliability tubes include microminiature metal-ceramic tubes, called *nuvistors*. Their metal-ceramic envelopes are 11 mm in diameter and 20 to 25 mm in height. The cylindrical electrode structure is mounted on a ceramic base wafer with the help of flanges

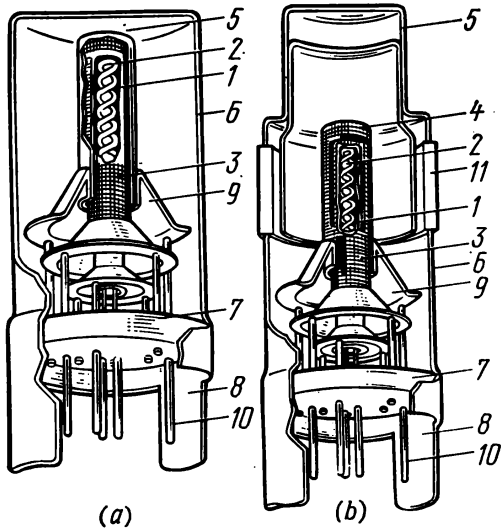


Fig. 4-29. Construction of nuvistors

(*a*) triode; (*b*) tetrode; *1*—cathode; *2*—heater; *3*—first grid; *4*—second grid; *5*—plate; *6*—metal envelope; *7*—ceramic base wafer; *8*—alignment ribs; *9*—washer; *10*—pins

(Fig. 4-29). Owing to such construction, nuvistors can be manufactured automatically with a very small spread in parameters.

Nuvistors show an increased resistance to mechanical factors and can operate at a temperature of —60° to +200°C and a relative humidity of up to 98%.

There also exist high-reliability tubes with a service life of up to 10,000 hours. Here, special attention is given to the service life of the cathodes and electrode supports which are made of very pure metals. These are tubes with a very high vacuum.

Electrometer Tubes. These are used in devices intended to measure extremely small currents (10^{-9} to 10^{-15} A). Here, the grid current is reduced to 10^{-8}—10^{-14} A. Electrometer tubes are manufactured as triodes, screened-cathode tetrodes, etc. To reduce leakage current, the control grid is mounted on ceramic or quartz spacers. These tubes also have getters to clean up any remaining gases.

Electrometer tubes operate at a very low plate voltage (5 to 8 V); their plate current does not exceed 300 μA; the transconductance is also very low, being 20 to 80 μA/V.

Chapter Five

OPERATIONAL HINTS FOR VACUUM TUBES

5-1. TUBE NOISE

Noise Effect. When discussing the operation of various vacuum tubes, we assumed that filament voltage was maintained constant, tube electrodes were energized by high-stability direct voltage sources, and plate current did not vary with time. In practice, however, the number of electrons reaching the plate per unit time varies about some average value. These variations, known as *electric fluctuations*, are very small as compared with the average current. For example, in receiving and amplifying tubes whose average plate current is about ten milliamperes, fluctuations are approximately one ten-millionth of $I_{p,av}$. If, however, such a tube serves as the first stage of a high-gain multitube amplifier, fluctuations at the amplifier output may reach several milliamperes and cause an appreciable fluctuation voltage across the plate load.

If the signal to be amplified is comparable with the fluctuations in the first amplifier tube, the output signal will be masked by the fluctuation voltage. Because in the amplifier load, such as headphones, random fluctuations are heard as hissing or rusning noise, the term *noise* has come to describe any type of fluctuations.

The smallest input signal that will produce a discernible signal at the output of the amplifier (or receiver), referred to as the *minimum discernible signal*, depends on the level of amplifier (or receiver) noise. Thus, tube noise makes it impossible to build a highly sensitive radio receiver that would be able to receive infinitesimal signals.

Sources of Tube Noise. Tube noise can arise from a large number of causes.

Random emission and arrival of electrons at the plate of tube brings about what is known as *shot noise**. The number of electrons,

* Shot noise derives its name from the resemblance that the noise bears to that of fine shot falling upon a diaphragm (also known as Schrot and Schottky noise).— *Translator's note.*

n, reaching the plate (provided plate current is equal to emission current) varies with time about some average value

$$\bar{n} = I_e \, \Delta t/e \tag{5-1}$$

where I_e is the emission current, e is the electron charge, and Δt is the time interval over which n is averaged.

If n electrons arrive at the plate during the time Δt, the deviation in plate current from the average value during the same time is

$$i_{\Delta t} = e\,(n - \bar{n})/\Delta t \tag{5-2}$$

The fluctuation component of plate current can be imagined as a series of very short current pulses produced by one or several "extra" electrons hitting the plate. The frequency of these pulses varies from subaudio to ultrahigh. The plate circuit of a tube, even if it contains a low purely resistive load, has a limited, though considerable, bandwidth Δf which depends on C_{pk}, load resistance, R_L, and some other factors. Therefore, only a part, Δf, of the fluctuation spectrum which ideally is infinitely wide, appears in the plate circuit of a tube.

Applying the method of harmonic (Fourier) analysis to fluctuating quantities and recalling Eq. (5-2), we may define the mean-square noise current with the help of an equation given in reference [10] as

$$\overline{i_n^2} = 2eI_p\Delta f \tag{5-3}$$

We have written this equation on the asumption that I_p is equal to I_e, so it holds only for the condition of saturation. When the emission of a tube is space-charge limited, there is a slight decrease in shot noise. If, at a certain instant, the cathode emits a larger number of electrons, the space charge will become denser and, as a consequence, the potential barrier will grow higher. A smaller number of electrons will be able to overcome it, and the plate current will increase by an amount smaller than the increment in the emission current. Thus, the space charge smoothes variations in the emission current as it were, and reduces noise. This effect is accounted for by a factor, Γ^2, in the expression for the mean-square fluctuation given in reference [10]

$$\overline{i_n^2} = 2e\Gamma^2 I_p\Delta f \tag{5-4}$$

The factor Γ^2 is known as the *space-charge noise-reduction factor* and defined as

$$\Gamma^2 = 0.644\,(2kT_k g_m)/eI_p \tag{5-5}$$

where k is Boltzmann's constant, T_k is the absolute temperature of the cathode, and g_m is the transconductance of a diode.

At $T_k = 1000$ K

$$\Gamma^2 = (\text{approx.})\ 0.11\ g_m/I_p \tag{5-6}$$

Equation (5-4) is in good agreement with experimental data.

Shot noise in triodes can also be described by Eq. (5-4), the numerator of Eq. (5-6) should be the transconductance of the equivalent diode rather than that of the triode

$$g_{md} = \frac{g_m}{\partial V_d / \partial V_g} = g_m/\delta \tag{5-7}$$

Multigrid tubes are more noisy than diodes and triodes. This is because, in addition to shot noise, there is a random division of current between electrodes, which gives rise to what is known as *partition noise*. The mean-square partition noise current is defined by an equation given in reference [10]

$$\overline{i^2_{p,\,cp}} = 2e \frac{k_{cp}}{k_{cp}+1} (\Gamma^2 I_2 + I_{g2}) \Delta f \tag{5-8}$$

Here $k_{cp} = I_p/I_{g2}$ is the current partition ratio.

Another source of tube noise is the random emission of secondary electrons from the tube electrodes. The secondary emission ratio varies with time and, consequently, there are fluctuations in secondary emission current, which raise total fluctuations. The mean-square noise current due to secondary emission is defined by an equation given in reference [10]

$$\overline{i^2_{p,\,\sigma}} = 2eI_p\sigma (\Gamma^2\sigma + 1) \Delta f \tag{5-9}$$

Total plate current fluctuations in multigrid tubes, combining partition and secondary-emission fluctuations, may be defined as

$$\overline{i^2_{p,\,n}} = 2e \left[\frac{k_{cp}}{k_{cp}+1} (\Gamma^2 I_p + I_{g2}) + I_p\sigma (\Gamma^2\sigma + 1) \right] \Delta f \tag{5-10}$$

In conventional receiving and amplifying tubes, the noise caused by secondary emission is usually much lower than shot or partition noise. So, in practice, the second term in Eq. (5-10) may be neglected.

Noise may be caused by the *flicker effect* which results from random variations in the work function of various cathode spots. The frequency of flicker noise is mainly in the audio and subaudio regions.

There is also ionization noise which is due to ionization by collision.

External mechanical factors may also contribute to tube noise, because as tube elements vibrate the distance between the electrodes changes and, as a consequence, so does plate current.

Calculation of Noise Performance. It is convenient to describe tube noise in terms of the *equivalent noise resistance*, because current fluctuations in a tube may be compared to fluctuations in an ordinary resistor.

As will be recalled, thermal agitation causes the random motion of electrons within any body. If we apply an external potential

difference to a body containing a number of free electrons and offering a certain electric resistance, R, the velocity vector of these electrons will have a component opposing the electric field strength vector, $\mathscr{E}$. Nevertheless, the electrons will continue to move at random, because their motion depends only on the temperature of the body. Therefore, an unequal number of electrons may pass through a cross section of the body per unit time in either direction. This gives rise to fluctuations in the current flowing in the current flowing in the circuit. The mean-square thermal-agitation noise voltage across a resistor R is defined by an equation given in reference [10], known as the Nyquist equation

$$\overline{v_n^2} = 4kTR\,\Delta f \tag{5-11}$$

where T is the absolute resistor temperature, and the other symbols have been defined before.

Equation (5-11) indicates that $\overline{v_n^2}$ depends on the resistance and temperature of the resistor. As the temperature rises, the thermal

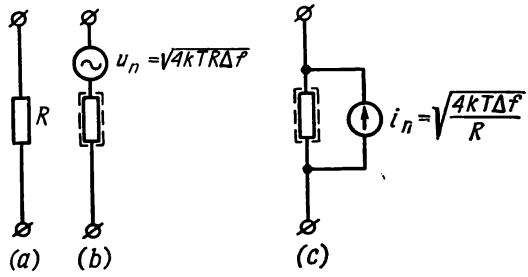


Fig. 5-1. Equivalent circuits of a noisy resistor
(*a*) noisy resistor; (*b*) noise-free resistance and equivalent noise voltage generator; (*c*) noise-free resistance and equivalent noise current generator

agitation of electrons increases. On the basis of Eq. (5-11), we can derive equations for the mean-square noise current and the available noise power across R thus

$$\overline{i_n^2} = \overline{v_n^2}/R^2 = 4kT\frac{1}{R}\Delta f \tag{5-12}$$

and

$$P_{na} = \frac{1}{4}\frac{\overline{v_n^2}}{R} = kT\,\Delta f \tag{5-13}$$

The "noisy" resistor R (Fig. 5-1*a*) may be represented by one of two equivalent circuits. In one, it is replaced by a noise-voltage generator in series with a noise-free resistor of the same value as the actual noisy resistor (Fig. 5-1*b*); in the other, it is replaced by a noise-current generator in parallel with a noise-free resistor of the same value (Fig. 5-1*c*).

When using the concept of an equivalent noise resistance, plate current fluctuations are calculated relative to the input of a tube. Here, the noisy tube (Fig. 5-2*a*) is replaced by a noise-free tube whose input circuit contains an equivalent noise voltage generator producing the same fluctuations in the plate circuit (Fig. 5-2*b*). For the triode, the calculation can easily be carried out with the well-known relation, $dI_p = g_m\, dV_g$. The equivalent noise voltage generator may in turn be replaced by a noisy resistor R_n at a temperature

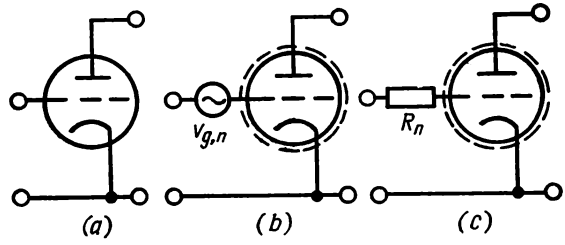


Fig. 5-2. Equivalent circuits of a noisy tube
(*a*) noisy triode; (*b*) noise-free triode and equivalent noise voltage generator; (*c*) noise-free triode and equivalent noisy resistor

T (usually at room temperature), which develops the same noise voltage (Fig. 5-2*c*).

Using Eq. (5-4) and recalling Eqs. (5-5) and (5-7), we can derive an equation for the mean-square noise voltage in the triode

$$\overline{v_{gn}^2} = \overline{i_{p,\,n}^2}/g_m^2 = \frac{0.644 \times 4kT_k}{g_m \delta}\,\Delta f \tag{5-14}$$

Equating (5-14) and (5-11), we get an equation for the equivalent noise resistance of the triode

$$R_{n,\,tr} = 0\ 644T_k/g_m T\delta \tag{5-15}$$

In Eq. (5-7) for the diode, $\delta = 1$, so

$$R_{n,\,d} = 0.644T_k/g_m T \tag{5-16}$$

The equivalent noise resistance for the pentode can be derived from Eq. (5-8)

$$R_{n,\,p} = \frac{k_{cp}}{k_{cp}+1}\,(0.644T_k/g_m T\sigma + 20I_{g2}/g_m^2) \tag{5-17}$$

Oxide-coated cathodes are usually operated at T_k = (approx.) $\times$ 1000 K. Substituting this value and $\delta = 0.8$ into Eqs. (5-15) through (5-17) and taking the temperature of the equivalent noise resistor to be $T = 300$ K, we get equations which can conveniently be used in practice.

Tube type	Equivalent noise resistance
Diode	$R_{n,d} = (\text{approx.})\ 2.3/g_m$
Triode	$R_{n,tr} = (\text{approx.})\ 2.5/g_m$
Pentode	$R_{n,p} = (\text{approx.})\ \frac{k_{cp}}{k_{cp}+1}(2.5/g_m + 20I_{g2}/g_m^2)$

The value of R_n ranges from 200 ohms for miniature triodes to hundreds of kilohms for frequency-converter tubes.

The noise performance of a tube may also be calculated using the concept of the relative noise temperature. By this method, the noise power developed by a two-port network, which may be a vacuum tube, is compared with the thermal noise power developed in an arbitrary linear circuit held at room temperature.

Assume that at room temperature, T, the total noise power due to all sources in a tube is $P_{n,t}$. The noise power developed in a resistor at room temperature is defined by Eq. (5-13). In the general case, $P_{n,t} \neq kT\ \Delta f$. The temperature of the resistor, however, may be adjusted to a value, T_{eq} at which $P_{n,t} = kT_{eq}\ \Delta f$.

The relation

$$t = T_{eq}/T \tag{5-18}$$

is known as the *relative noise temperature.* For example, recalling Eqs. (5-4) and (5-5), we may write for a diode with the dynamic plate resistance $R_p = 1/g_m$

$$t_d = 0.644T_k/T = (\text{approx.})\ 2.3$$

The noise performance of two-port networks including circuits built around vacuum tubes or other electron devices is described in terms of the *noise factor.* The noise factor (also called the *noise figure*), designated F, is the ratio of the total noise power at the output of a two-port to the output noise power due to the thermal noise of a signal source

$$F = P_{n,\,out,\,tot}/P_{n,\,out,\,s} \tag{5-19}$$

The noise factor is a dimensionless quantity, although sometimes it may be expressed in decibells

$$F = 10 \log_{10} P_{n,\,out,\,tot}/P_{n,\,out,\,s} \tag{5-20}$$

Reduction of Tube Noise. From the foregoing, it follows that tube noise is directly proportional to the cathode temperature T_k and inversely proportional to the transconductance g_m. In addition, tube noise increases with rising current partition ratio k_{cp}, screen grid current I_{g2} and secondary emission ratio.

The cathode temperature can hardly be reduced significantly because this would affect other tube parameters and ratings. In special low-noise tubes, however, one still seeks to reduce T_k. The value of

R_n can markedly be reduced by raising the transconductance of a tube. This is achieved by placing the control grid as close as possible to the cathode and reducing the grid-wire spacing. This, however, cannot always be done without affecting other tube parameters.

As can be seen from Eq. (5-17), noise in multigrid tubes may be diminished by reducing the current partition ratio k_{cp} (for which purpose, the screen grid wires are spaced wide apart) or screen-grid voltage V_{g2} which, in turn, scales down I_{g2}. The secondary emission ratio can be reduced either by using tube electrodes made from metals with a low secondary emission ratio or by giving them a coat of low-σ metal.

5-2. EQUIVALENT CIRCUITS OF VACUUM TUBES

When calculating devices containing resistors, capacitors and inductances in addition to vacuum tubes, it is very convenient to replace vacuum tubes by equivalent circuits.

Equivalent Circuit for the Vacuum Diode. In the general case, two voltages may be applied between the cathode and plate of a

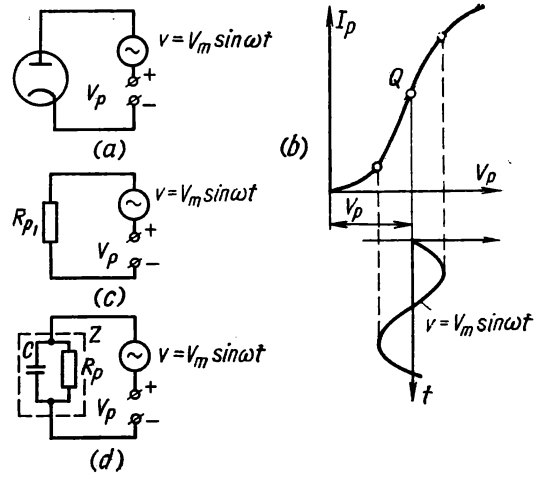


Fig. 5-3. Equivalent circuits of the vacuum diode

(*a*) general connection of the diode in a circuit; (*b*) portion of the diode characteristic approximated by a straight-line segment; (*c*) equivalent circuit of the diode for low frequencies; (*d*) equivalent circuit of the diode for high frequencies

vacuum diode. One is the direct voltage, V_p, which positions the quiescent point on the diode volt-ampere characteristic, and the other is an alternating voltage, $v = V_m \sin \omega t$ (Fig. 5-3*a*). If the amplitude of the alternating voltage is not very high, the operating region of the characteristic around the quiescent point, Q (Fig. 5-3*b*), may be considered linear. Accordingly, the diode may be replaced by a resistor R_p equal to its dynamic plate resistance at this point. The interelectrode capacitance C should be connected in parallel

with R_p. At audio frequencies, however, where the capacitive reactance $X_C = 1/i\omega C$ considerably exceeds R_p, the shunting effect of X_C may be neglected, and the equivalent circuit for the vacuum diode will be as one shown in Fig. 5-3*c*.

If the frequency of the a.c. signal, ω, is high, the capacitive reactance may not be ignored, and the diode should be treated as an impedance Z (Fig. 5-3*d*).

Equivalent Circuits for the Triode. As has been shown in Chapter 3, the vacuum triode may be represented as a linear resistive two-port whose output (plate) and input (grid) currents are defined by Eqs. (3-24) and (3-25), respectively

$$dI_p = (\partial I_p/\partial V_g)\, dV_g + (\partial I_p/\partial V_p)\, dV_p$$

$$dI_g = (\partial I_g/\partial V_g)\, dV_g + (\partial I_g/\partial V_p)\, dV_p$$

Here, the partial derivatives (the characteristic admittances of the tube) serve as the static (d.c.) parameters. On the basis of Eqs. (3-33), (3-36), (3-45) and (3-47), we may re-write the above equations as

$$dI_g = g_g\, dV_g + g_{gp}\, dV_p \tag{5-21}$$

$$dI_p = g_m\, dV_g + \frac{1}{R_p}\, dV_p \tag{5-22}$$

For small-signal operation when the path of the Q-point on the characteristic may safely be replaced by a straight line, the incremental changes in voltages and currents in Eqs. (5-21) and (5-22) may be replaced by their instantaneous values

$$i_g = g_g v_g + g_{gp} v_p \tag{5-23}$$

$$i_p = g_m v_g + \frac{1}{R_p}\, v_p \tag{5-24}$$

The equivalent circuit for the vacuum triode based on Eqs. (5-23) and (5-24) is shown in Fig. 5-4*a*. If the grid draws no current ($i_g = 0$), the equivalent circuit given in Fig. 5-4*a* reduces to that for the triode plate circuit containing a $g_m v_g$-current generator (Fig. 5-4*b*).

olving Eq. (5-24) for v_p and recalling that $g_m R_p = \mu$, we get

$$v_p = R_p i_p - \mu v_g \tag{5-25}$$

This equation is represented by the equivalent plate circuit containing a constant-voltage generator (Fig. 5-4*c*).

For the triode containing a load resistor R in the plate circuit, the circuits given in Fig. 5-4*b* and *c* take the form of those of Fig. 5-4*d* and *e*.

The foregoing holds not only for the vacuum triode. If a tetrode or pentode is used as an amplifier, the screen- and suppressor-grid

circuits are grounded for the alternating component. The suppressor grid is usually connected to the cathode and returned to ground via a high-value capacitor shunting the cathode bias resistor. As a rule, the screen grid is also returned to ground through a high-value capacitor. So, no alternating voltages are developed across these circuits. The input of a tetrode or pentode amplifier, as well as that of a triode amplifier, is the control-grid circuit. The amplified

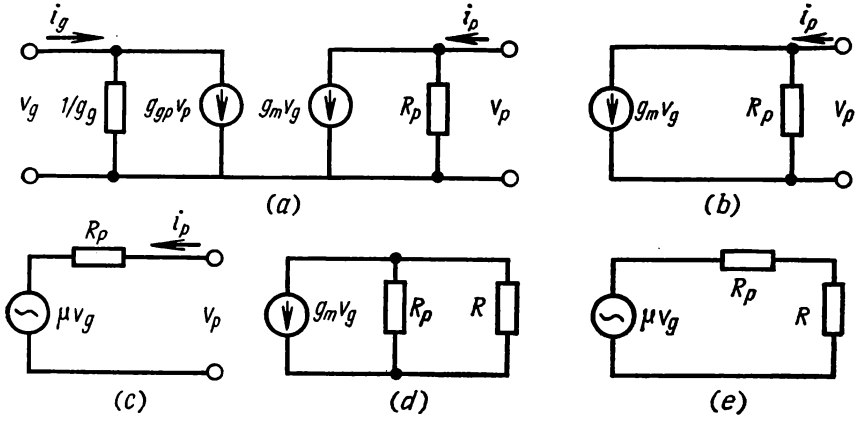


Fig. 5-4. Equivalent circuits of the vacuum triode for low frequencies
(*a*) equivalent circuit of a two-port; (*b*) equivalent plate circuit containing a current generator; (*c*) equivalent plate circuit containing a voltage generator; (*d*) and (*e*) equivalent plate circuits with a load resistor

voltage is taken from the plate load and, as a consequence, the plate circuit is the output. Thus, the circuits given in Fig. 5-4 may also serve as equivalent circuits for the tetrode and pentode.

Equivalent Circuits for High Frequencies. In Sec. 4-6 we have learned that when a tube operates at high frequencies, the characteristic admittances become complex due to the transit-time effects and distributed reactances. So, Eqs. (5-23) and (5-24) should be re-written as follows

$$\dot{I}_g = Y_{in}\dot{V}_g + Y_{tr}\dot{V}_p \tag{5-26}$$

$$\dot{I}_p = \dot{g}_m\dot{V}_g + Y_{out}\dot{V}_p \tag{5-27}$$

The input admittance of a tube is

$$Y_{in} = \dot{I}_g/\dot{V}_g = G_{in} + i\omega C_{in} \tag{5-28}$$

where $G_{in} = G_{in\tau} + G_{inL}$.

The quantities $G_{in\tau}$ and G_{inL} are defined by Eqs. (4-42) and (4-44). The input capacitance of the triode is $C_{in} = C_{gk}$; that of the

pentode, according to Eq. (4-44), is

$$C_{in} = (\text{approx.})\, C_{g1k} + C_{g1g2}$$

The transfer admittance depends on the capacitance between the plate and control grid

$$Y_{tr} = \dot{I}_g / \dot{V}_g = i\omega C_{pg1} \tag{5-29}$$

The transconductance

$$g_m = \dot{I}_p / \dot{V}_g \tag{5-30}$$

as has been shown in Sec. 4-6, is also a complex quantity.

The output admittance is

$$Y_{out} = \dot{I}_p / \dot{V}_p = 1/R_p + i\omega C_{out} \tag{5-31}$$

Here, the output capacitance, C_{out}, for the triode is equal to C_{pk}, and that for the pentode is defined by Eq. (4-34)

$$C_{out} = (\text{approx.})\, C_{pk} + C_{pg3} + C_{pg2}$$

The equivalent circuit for high frequencies developed on the basis of Eqs. (5-26) and (5-27) is shown in Fig. 5-5*a*. If we ignore the effect that plate voltage has on grid current and deem that $\dot{I}_g$ depends only

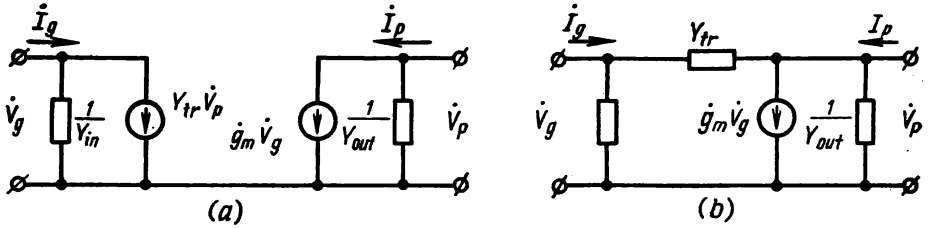


Fig. 5-5. Equivalent circuits of the vacuum triode for high frequencies (*a*) equivalent circuit of a two-port; (*b*) equivalent circuit containing one current generator

on the input admittance, this circuit will take the form shown in Fig. 5-5*b*.

5-3. APPROXIMATION OF TUBE CHARACTERISTICS

We have learned that the static (d.c.) characteristics of vacuum tubes are nonlinear. So, in the general case, a vacuum tube is a nonlinear element which does not obey Ohm's law.

Methods of Approximation. Attempts to describe the relation $I_p = f(V_p)$ for the vacuum diode in mathematical form have led to the three-halves power law. However, it fits the empirical curve very approximately and then only within a limited range of the argument. A more precise mathematical description of the plate characteristic

has not been found yet, nor can one expect this to be a simple law convenient for practical calculations. So, in practice, tube characteristics are approximated by analytical functions as simple as possible but sufficiently accurate in describing the empirical relations (this is known as *nonlinear approximation*). The approximating functions are power polynomials, trigonometrical functions, exponential functions, etc.

In practical calculations, use is widely made of piecewise-linear approximation in which a tube characteristic is replaced by a broken line made up of one or several straight-line segments.

Piecewise-linear approximation is applicable to small signals when the error caused by the replacement of the operating region by a straight line is insignificant.

Piecewise-Linear Approximation. As already mentioned, this technique consists in replacing an empirical characteristic curve by several straight-line segments. The greater the number of straight-line segments, the closer the approximation. In practice, characteristics are usually divided into three to five segments, so the approximation is rather rough. The simplicity of analysis, however, compensates for the errors inevitable in this technique.

Let us discuss piecewise-linear approximation as applied to the static (d.c.) characteristics of a tube operating at audio frequencies when the interelectrode capacitances are high and their effect may be neglected.

Approximation of Diode Characteristics. Figure 5-6*a* shows the plate characteristic of the vacuum diode, $I_p = f(V_p)$. According to the manner in which the tube is operated, we can distinguish several portions on the characteristic to be approximated. The simplest case is when the diode operates as a rectifier and plate current flows in the tube only when the plate is at a positive potential

$$\left.\begin{array}{ll} \text{at } V_p \leqslant 0 & I_p = 0 \\ \text{at } V_p > 0 & I_p > 0 \end{array}\right\} \qquad (5\text{-}32)$$

The approximate characteristic is shown in Fig. 5-6*b* and the equivalent circuit for the diode appears in Fig. 5-6*c*. The "1" position of switch S corresponds to region *1* on the characteristic: the circuit is open and the plate current is zero. The "2" position of the switch corresponds to region *2*. The slope of the straight line drawn through the origin is determined by the average dynamic plate resistance R_p over the section OA.

Under certain conditions, the tail of the real diode characteristic may be immaterial due to the low value of current. Then, it is preferable to approximate the curve shown in Fig. 5-6*d* by the straight-line segment which makes the intercept V_2 on the x-axis (Fig. 5-6*e*). Here, the slope is naturally different, because it is determined by the

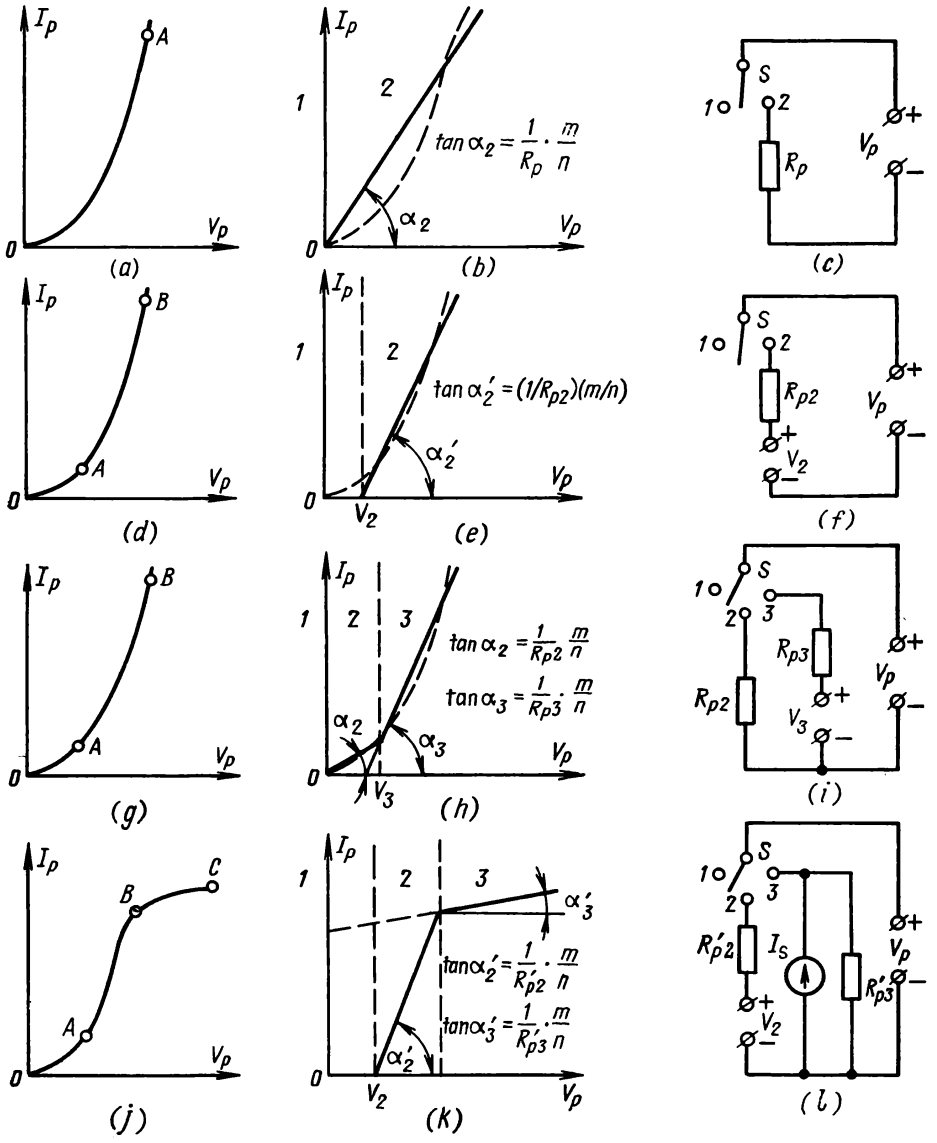


Fig. 5-6. Piecewise-linear approximation of the diode plate characteristic (*a*), (*d*), (g) and (*j*) characteristic portions to be approximated; (*b*), (*e*), (*h*) and (*k*) approximated characteristics; (*c*), (*f*), (*i*) and (*l*) equivalent circuits of the diode; m and n are the scale factors

average dynamic plate resistance over the section AB. In addition to resistance R_{p2}, the equivalent circuit of Fig. 5-6*f* should contain a direct-voltage source, V_2, because

$$\left.\begin{array}{ll} \text{at } V_p \leqslant V_2 & I_p = 0 \\ \text{at } V_p > V_2 & I_p > 0 \end{array}\right\} \qquad (5\text{-}33)$$

One more case where the tail of the characteristic is of importance is illustrated by Fig. 5-6*g* through *i*. Figure 5-6*j* through *l* shows the approximation of the characteristic under conditions of saturation. In contrast to the previous case, the tube operating within region *3* is represented in the equivalent circuit by an equivalent current generator, I_s, connected in parallel with resistance R'_{p3}. This is more

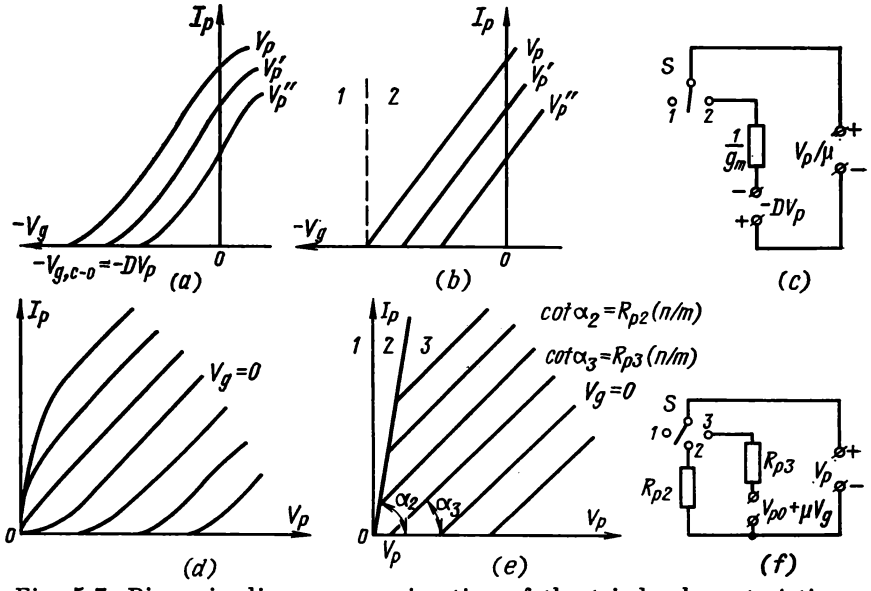


Fig. 5-7. Piecewise-linear approximation of the triode characteristics
(*a*) family of plate-grid characteristics; (*b*) approximated plate-grid characteristics; (c) equivalent circuit for this approximation; (*d*) family of plate characteristics; (*e*) approximated plate characteristics; (*f*) equivalent circuit for this characteristic; *m* and *n* are the scale factors

convenient because within region *3* the straight line intersects the voltage axis at a very negative V_p.

Approximation of Triode Characteristics. Figure 5-7 illustrates approximation of the static (d.c.) characteristics of the triode. A family of plate-grid triode characteristics (Fig. 5-7*a*) is approximated by parallel straight lines whose slope is determined by the average slope between $V_g = V_{g,c-o}$ and $V_g = 0$ (Fig. 5-7*b*). The equivalent circuit of Fig. 5-7*c* contains a resistance, $1/g_m$, and a voltage generator, $-DV_p$, determining the cut-off voltage of the tube and, as a consequence, locating the point at which the characteristic intersects the x-axis. Each curve corresponds to a particular value of V_p and also to a particular emf due to the equivalent voltage generator. Region *2* in the plate characteristics (Fig. 5-7*d*), corresponding to the condition of electron flyback at $V_g > 0$, is a nonoperating region. In the equivalent circuit of Fig. 5-7*f*, it is represented by a resistance R_{p2}. The operating region, *3*, is represented in the equivalent cir-

cuit by a resistance R_{p3} (on the assumption that the R_p of the tube remains the same with varying V_g and V_p) and a generator whose voltage at $V_g = 0$ is equal to V_{p0}. As V_g takes on a new value (which corresponds to a different characteristic), the cut-off voltage of the tube changes by μV_g.

Approximation of Tetrode and Pentode Characteristics. The plate characteristic of the tetrode (Fig. 5-8*a*) can be approximated by three

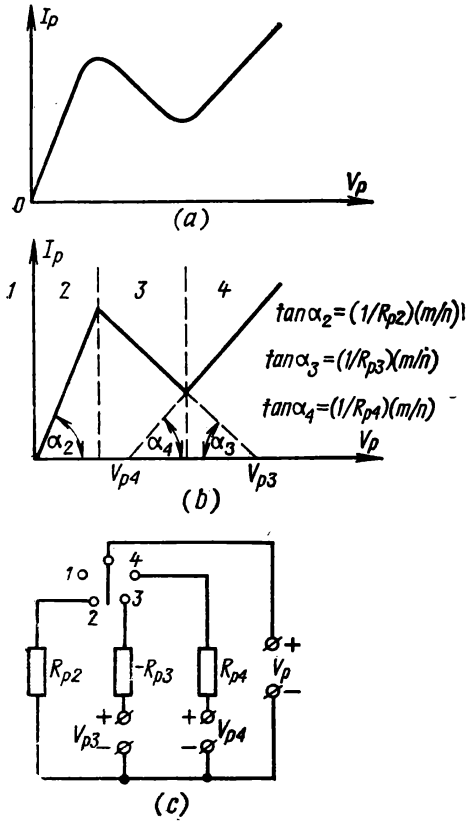


Fig. 5-8. Piecewise-linear approximation of the tetrode plate characteristic

(*a*) plate characteristic; (*b*) approximated plate characteristic; (*c*) equivalent circuit; *m* and *n* are the scale factors

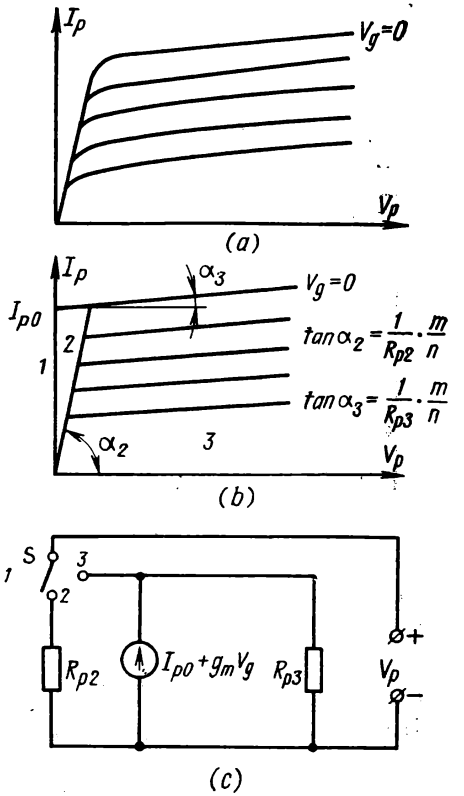


Fig. 5-9. Piecewise-linear approximation of the pentode characteristics

(*a*) plate characteristics; (*b*) approximated plate characteristics; (*c*) equivalent circuit; *m* and *n* are the scale factors

straightline segments *2*, *3* and *4* (Fig. 5-8*b*). Here, the main distinction is the dynatron effect in region *3*, which is represented in the equivalent circuit by a negative resistance $-R_{p3}$ and a direct voltage generator, V_{p3}.

As with the diode at saturation, the operating region, *3*, (Fig. 5-9*b*) on the plate characteristics of the pentode (Fig. 5-9*d*) can convenient-

ly be represented by an equivalent circuit containing a current generator (Fig. 5-9*c*).

A.C. (With Signal) Equivalent Circuit. If we apply a signal, that is, an alternating voltage, $v_g = V_{gm} \sin \omega t$, to the input of a tube, the equivalent circuit for this tube should contain an equivalent alternating voltage or current generator in addition to direct voltage sources. Figure 5-10 shows the triode equivalent circuit of Fig. 5-7*f* extended to include generators μv_g or $g_m v_g$ so as to illustrate operation of an amplifier.

5-4. RELIABILITY OF VACUUM TUBES. OPERATING CONDITIONS AND RATINGS

Reliability is a concept which is widely used to evaluate the ability of a device to operate without failure for a certain period of time. Quantitatively, reliability is expressed in terms of several criteria.

One of these criteria is the *failure rate.*

The failure rate is generally the rate at which failure occurs during an interval of time as a function of the total interval length

$$\lambda(t) = (\text{approx.})\ \Delta d_i / (n - d)\ \Delta t_i \tag{5-34}$$

Here, n is the total number of trials, d_i is the number of failures during Δt_i, and d is the number of failures before Δt_i.

Another common measure of reliability is the *probability of survival*

$$P(t) = \lim_{\substack{n\to\infty \\ \Delta t\to 0}} \frac{n - \sum_{i=1}^{t/\Delta t} \Delta d_i}{n} = (\text{approx.}) \frac{n - \sum_{i=1}^{t/\Delta t} \Delta d_i}{n} \tag{5-35}$$

These criteria are related as follows

$$P(t) = \exp\left(-\int_0^t \lambda(t)\, dt\right) \tag{5-36}$$

The failure rate is the most common measure of reliability. Graphically, the failure rate of vacuum tubes is represented as shown in Fig. 5-11. The initial portion of the curve* represents the *shake-down* or *wear-in period* during which failures occur mainly due to hidden manufacturing defects. The middle portion represents the constant failure-rate or useful-life period during which failures due to various factors occur before the end of the average service life, t_2. The end portion represents the wear-out period during which the failure rate rapidly grows.

Vacuum tubes are prone to several kinds of failures. These are *complete (sudden) failures* caused by broken leads, electrode short-

* Sometimes called the "bath-tub" curve.— *Translator's note.*

circuits, cracks in envelopes, etc., and *partial failures* taking place when at least one rating fails to meet the specifications. These are generally mechanical failures resulting from wear-out.

In Chapter 4, we have considered some design and manufacturing measures which yield high-reliability and long-life vacuum tubes. Such tubes have a failure rate as low as λ = (approx.) 0.3×10^{-5} to 4×10^{-5} per 500 hours (which is one-half to one-sixth that of conventional tubes). The failure rate of vacuum tubes rises when

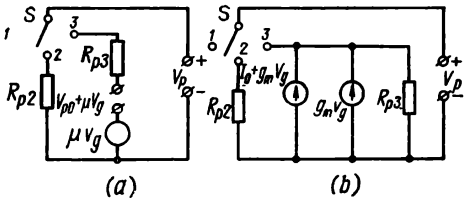


Fig. 5-10. A.C. equivalent circuit for the triode
(*a*) equivalent circuit containing voltage generators; (*b*) equivalent circuit containing current generators

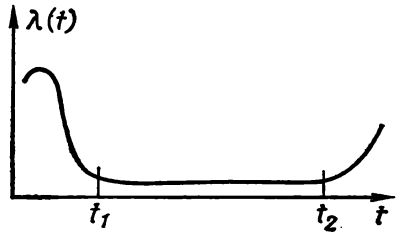


Fig. 5-11. Failure rate as a function of time

they operate near the limit of service conditions, at an increased ambient temperature, under conditions of vibration, shocks, etc.

Tube Ratings. In addition to nominal and bogey values, tube manufacturers specify ratings which establish either *limiting capabilities* or *limiting conditions* for a tube. These include maximum and minimum filament voltage, maximum plate and screen-grid voltage, maximum plate and screen-grid dissipation, and so on.

A circuit based on a vacuum tube should be designed so that the tube ratings will not be exceeded.

Mechanical, Thermal and Climatic Operating Conditions. Data sheets also specify mechanical and climatic operating conditions including maximum acceleration under conditions of vibration at various frequencies, maximum acceleration under conditions of multiple and single shocks, ambient temperature and relative humidity. For some tubes, data sheets give the maximum safe temperature for the envelope and the operating position in apparatus (for example, any or only vertical).

Spread in Characteristics and Parameters. When calculating circuits built around vacuum tubes, it is necessary to take into consideration that data sheets specify average (bogey) tube characteristics and parameters. Because of inevitable variations in tube manufacture, characteristics and parameters may vary considerably from tube to tube. For example, two tubes of the same type having equal grid and plate voltages may differ in plate current by 20 to 30%; the difference in grid leakage current may reach several hundred times.

Chapter Six

PHOTOELECTRIC DEVICES

6-1. GENERAL

Definition. A *phototube* (*or photocell*) is a vacuum, gas-filled or semiconductor device which converts the energy of electromagnetic radiation into electric signals.

The term phototube applies to electron devices responding to visible electromagnetic radiation ($\lambda \approx 0.38$ to 0.78 μm), the near infrared ($\lambda \approx 0.78$ to 340 μm) and ultraviolet ($\lambda \approx 10^{-2}$ to 0.38 μm).

Principle of Operation. For their operation, phototubes depend on the internal electric processes caused by absorption of electromagnetic radiation. When a solid specimen is exposed to radiation, the incident radiant flux is partly reflected from and partly absorbed by the specimen. The quanta of the absorbed radiant flux change the energy of electrons in the solid. The electrons which have acquired an additional energy may escape from the solid (thus giving rise to photoelectric emission) or move to higher energy levels, say, from the valence to the conduction band. In the latter case, the number of mobile charge carriers increases and so does the conductivity of the material (this is known as the photoconductive effect). Also, electron-hole pair generation brought about by absorption of radiant energy may affect the processes near the *P-N* junction, leading to an increased current across the potential barrier or an additional potential difference.

Basic Classification. According to the processes stimulated by absorption of radiant energy, phototubes may be divided into two groups. One group includes *vacuum and gas-filled phototubes* which operate by the photoemissive effect; the other group includes *semiconductor photocells* whose operation is based on changes in the energy states of electrons in a solid.

The first group may further be classed into vacuum and gas-filled phototubes proper and *photomultiplier tubes.* It is these devices that are discussed in this chapter.

Semiconductor photocells may be divided into semiconductor photocells proper, *photoresistors*, *solar cells*, *photodiodes*, *phototran-*

sistors, etc. Semiconductor photoelectric devices will be discussed in Chapter 14.

A special group of *electron-beam photoelectric devices* are television camera tubes which combine an electron beam with the photoemissive and photoconductive effects. Because of their specific nature as regards operation and construction, television camera tubes will not be discussed in this text, and the reader should turn to a course on television.

In the Soviet Union, the diagram symbols for photoelectric devices are established by USSR State Standard GOST 7624-62. The nomenclature and definitions for photomultiplier tubes are set up by GOST 20526-75.

Radiant Energy. This term is often used to characterize the energy of electromagnetic radiation at optical wavelength which includes the visible, ultraviolet and infrared regions of the spectrum. Radiant energy, W, is usually measured in joules or electronvolts (1 J = $= 6.29 \times 10^{18}$ eV).

The time rate of flow of radiant energy is termed the *radiant flux*

$$F = dW/dt \tag{6-1}$$

It is expressed in watts (1 $W = 6.29 \times 10^{18}$ eV/s). The radiant flux perceived by the human eye is called the *luminous flux*, Φ. It is measured in lumens (lm). A standard luminous flux of one lumen is emitted by a hollow radiator (blackbody) with an emitting hole 5.3×10^{-3} cm^2 in area, held at the solidification temperature of platinum ($T = 2046$ K).

The sensitivity of the human eye varies with wavelength. It is maximal to radiation at $\lambda =$ (approx.) 0.555 μm (the green region of the spectrum). This sensitivity is taken to be unity ($K_{\lambda max} = 1$). The eye sensitivity to other wavelengths is expressed in terms of *relative sensitivity*, K_λ, which is a fraction of $K_{\lambda max}$.

A luminous flux of one lumen at $\lambda = 0.555$ μ is equal to 1/683 W.

With nonmonocromatic radiation, the luminous flux is determined in terms of the integral taken over the entire radiation spectrum

$$\Phi = 683 \int_0^\infty F(\lambda) K_\lambda(\lambda) d\lambda \tag{6-2}$$

The ratio of the luminous flux to the radiant flux is called the *luminous efficiency*; it is measured in lm/W

$$\psi - = \Phi/F = 683 \frac{\int_0^\infty F(\lambda) K_\lambda(\lambda) d\lambda}{\int_0^\infty F(\lambda) d\lambda} \tag{6-3}$$

The luminous flux in a given direction is expressed in terms of the *luminous intensity*

$$J = \Phi/\omega \tag{6-4}$$

The unit of luminous intensity is the candela which is defined as a luminous flux of one lumen per unit solid angle ($\omega = 1$ steradian).

6-2. PHOTOELECTRIC EMISSION

Definition. *Photoelectric emission* (or, simply, *photoemission*) is electron emission attributable solely to the electromagnetic radiation absorbed by a solid or liquid and not associated with its heating.

A cathode whose primary function is photoelectric emission is termed a *photoelectric cathode.*

Theory of Photoemission. According to the law derived by the Russian scientist A. G. Stoletov, the number of emitted electrons and, as a consequence, the saturation photocurrent, I_{ph}, are proportional to the incident luminous flux

$$I_{ph} = n_{ph}e = K\Phi \tag{6-5}$$

Here, n_{ph} is the number of photoelectrons emitted per unit time, Φ is the luminous flux, and K is the constant of proportionality.

In compliance with the Einstein photoelectric equation, the kinetic energy of an electron ejected from a system by an incident photon is a function of the frequency, ν, associated with the absorbed photon and is independent of the luminous flux intensity

$$mv^2/2 = h\nu - E_0 \tag{6-6}$$

Here, $h\nu$ is the energy of the absorbed photon and E_0 is the energy necessary to remove an electron from the system, called the *electronic work function* of that system.

The above two expressions are the basic laws of photoelectric emission.

Metals can emit electrons already at a temperature close to zero, because the incident photons can impart enough energy to the electrons near the Fermi level to escape from the surface of the material.

Photoemission in semiconductors is more complex. Here, the absorbed photons may impart their energy not only to the electrons in the conduction band, but also to impurity electrons and those in the valence band. At room temperature, the density of electrons in the conduction band is low, so the photocurrent established by these electrons is small. If the energy of an incident photon is sufficient for an electron to leave the impurity levels, photoemission increases. As the energy of the absorbed photons increases, more electrons from the valence band contribute to the major part of photocurrent.

Threshold of Photoemission. If all of the energy of the absorbed photon is expended to remove an electron from the material ($h\nu = E_0$), the kinetic energy of the emitted electron will be zero, If

$h\nu > E_0$, the electron, according to Eq. (6-6), will have some kinetic energy equal to the excess energy of the photon.

It is obvious that, as the frequency, ν, of a monochromatic radiation decreases, the energy of a photon also decreases until at a certain value $\nu = \nu_0$ the kinetic energy of the emitted electrons will be zero. The frequency $\nu_0 = 1/\lambda_0$ is known as the *threshold frequency*. As can easily be seen from Eq. (6-6), the threshold frequency is not the same for materials differing in work function. Thus, for each material, photoelectric emission starts at a certain threshold value and increases with increasing frequency ν so that electrons may be ejected from the lower energy levels. Photoemission at which photocurrent steadily rises with increasing frequency is called *normal*.

Impurities markedly affect photoelectric emission and also the threshold frequency. A rise in the impurity concentration brings about a fall in the threshold frequency.

Spectral Characteristic. It should be noted, however, that absorption of radiant energy expressed in terms of the *absorption factor* depends on the structure of the crystal lattice, impurities, surface conditions and some other factors, and varies with ν. Variations in the absorption factor naturally affect photocurrent. The relation between photocurrent, I_{ph}, and frequency, ν, at $\Phi = \text{const}$ is termed the *spectral characteristic of a given photocathode*. Photoelectric emission characterized by a very high photocathode sensitivity within a narrow wavelength interval is called *selective photoemission*.

External Field Effect. Photoemission conditions are markedly affected if a photocathode is subject to an external electric field. As with thermionic emission, an external electric field reduces and narrows the potential barrier, and this is accompanied by the *tunnel effect*. The effect of an accelerating external field is most pronounced near the threshold of photoemission where the velocities of a majority of photoelectrons are close to zero.

6-3. PHOTOTUBES

Definition. A *phototube* is an electron device whose electric properties are controlled by incident light.

According to the medium in which the photoeffect takes place, phototubes may be classed into *vacuum* and *gas-filled* types.

Construction. Figure 6-1 is a sketch of phototubes. The photocathode is made up by a layer of a light-sensitive material applied either to a part of the internal surface of the glass envelope or to a special plate. The most common are monatomic-film cathodes which consist of a thin layer of oxidized metal given a monatomic coat of cesium. Generally, use is made of *oxygen-cesium* and *antimony-cesium* photocathodes. In the majority of cases, the antimony-cesium photocathode is applied to a thin nickel base film.

The anode is either a ring placed in the middle of the envelope or a grid placed near the planar cathode. Such an anode does not prevent light from striking the photocathode.

The connection diagram of a phototube is shown in Fig. 6-2. The cathode is connected to the negative terminal and the anode to the positive terminal of a direct-voltage source, E. The anode circuit contains a load resistor, R_L. When light falls upon the tube, the

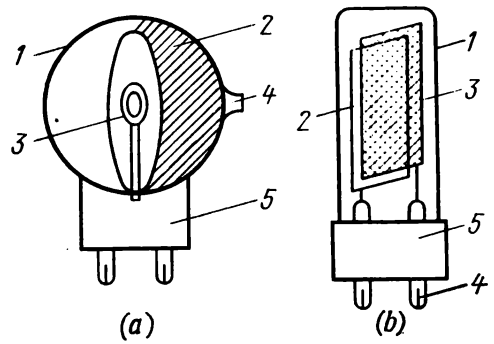


Fig. 6-1. Phototubes
(*a*) with the cathode on glass; (*b*) with a plate cathode; *1*—envelope; *2*—photocathode; *3*—anode; *4*—leads; *5*—base

photocathode emits electrons which are carried along by the positive anode field.

Volt-Ampere and Luminous Characteristics of Phototubes. The basic characteristics of phototubes are the volt-ampere characteristic, $I_{ph} = f(V_a)$ at $\Phi = \text{const}$, and the luminous characteristic, $I_{ph} = f(\Phi)$ at $V_a = \text{const}$.

Volt-ampere characteristics of a vacuum phototube are shown in Fig. 6-3*a*. At low V_a, not all of the electrons emitted by the photocathode reach the anode. Some of them are caught by a negative space charge near the cathode. As the anode becomes more positive, the space charge gradually disperses until all of the electrons emitted by the cathode can reach the anode. When this happens, saturation is said to take place. A further increase in anode current under this condition may be obtained only by raising photoemission.

The *luminous characteristic* of a vacuum phototube (Fig. 6-3*b*) directly relates its photocurrent to the luminous flux. The slope of the luminous characteristic is defined by the constant K in Eq. (6-5)

$$I_{ph} = K\Phi$$

The constant K is numerically equal to $\tan \alpha$ and is called the *integrated sensitivity*. It defines the photocurrent occurring when a phototube is exposed to a luminous flux of one lumen irrespective of its spectrum.

The volt-ampere and luminous characteristics of vacuum phototubes using antimony-cesium emitters without metal base layers

somewhat differ from those discussed above (Fig. 6-3*c* and *d*). Here, photocurrent is not directly proportional to the luminous flux over the entire range of Φ. This can be attributed to the semiconductor structure of the cathode. The point is that when the cathode is illumi-

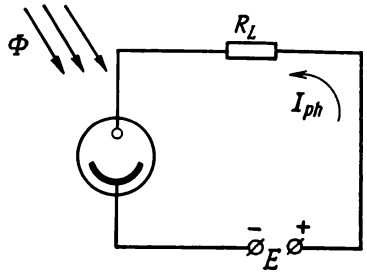


Fig. 6-2. Connection diagram of a phototube

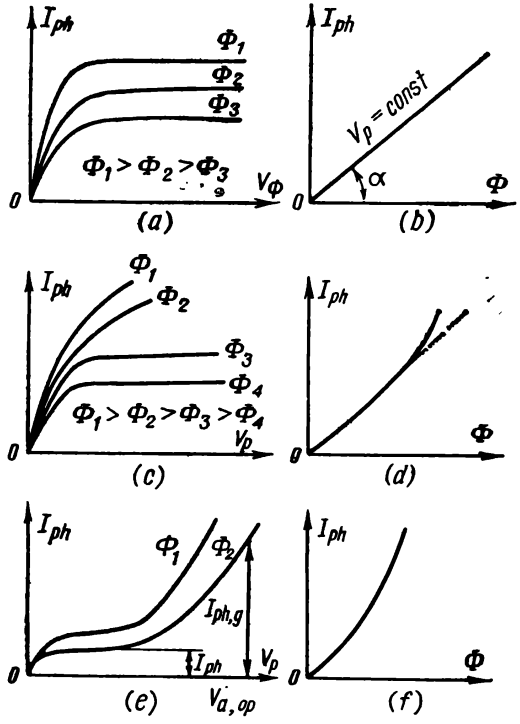


Fig. 6-3. Phototube characteristics

(*a*) volt-ampere characteristics of a vacuum phototube; (*b*) luminous characteristic of a vacuum phototube; (*c*) volt-ampere characteristics of a phototube using an antimony-cesium cathode on glass; (*d*) luminous characteristic of a phototube using an antimony-cesium cathode on glass; (*e*) volt-ampere characteristics of a gas phototube; (*f*) luminous characteristic of a gas phototube

nated and the resultant photocurrent flows along its surface towards the lead, the potential at the various points on the cathode rises with increasing distance from the lead and increasing photocurrent, Some electrons return to the positive areas of the cathode, thus causing secondary emission (Fig. 6-4). Because of secondary emission, the anode current is higher than in a phototube using a cathode with a metal base layer.

The characteristics of tubes using antimony-cesium cathodes with metal base layers do not differ from conventional tubes.

In actual photocathodes, however, photocurrent is non-zero at zero luminous flux; the so-called *dark current*, I_d, flows in a tube. This current is established by thermionic emission from the cathode and also by the current due to the conduction of the glass.

Gas-Filled Phototubes. The envelopes of gas phototubes are filled with a rarefied gas at a pressure of about a few tens of pascals. In these devices, electrons moving to the anode collide with gas molecules. At a certain kinetic energy, the electrons can ionize the gas

molecules. The newly formed electrons travel to the anode and the positive ions move to the cathode. This motion of electrons and ions raises the density of the charged-particle stream, and the anode current thus increases. Gas-filled phototubes utilize a dark discharge.

The volt-ampere and luminous characteristics of a gas-filled phototube are shown in Fig. 6-3*d* and *e*. When little light falls upon the tube, the number of photoelectrons is small and ionization is insignificant, so the photocurrent is directly proportional to the luminous flux. As Φ increases, more electrons collide with gas molecules and the anode current made up of electrons and ions rises. An increase in anode voltage brings about a rise in anode current for the same reason.

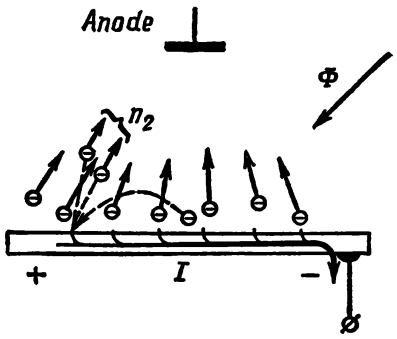


Fig. 6-4. Electron motion in a phototube using a cathode without a base layer

The increase in the photocurrent generated by gas phototubes as compared with the photocurrent in vacuum phototubes is customarily expressed in terms of the gas-amplification factor

$$K_g = I_{ph,\,g}/I_{ph} \qquad (6\text{-}7)$$

where $I_{ph\ g}$ is the current in gas-filled phototube at an operating anode voltage and a certain Φ, and I_{ph} is the saturation current in a vacuum phototube at the same luminous flux (Fig. 6-3*e*).

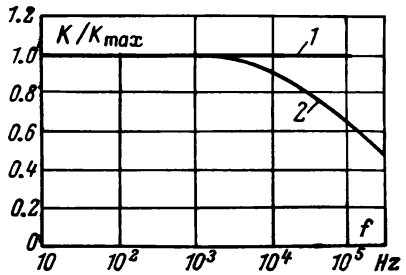


Fig. 6-5. Tube response as a function of photon frequency
1—vacuum phototube; *2*—gas phototube

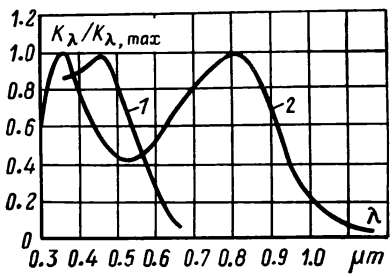


Fig. 6-6. Spectral characteristics of phototubes
1—with an antimony-cesium cathode and *2*—with an oxygen-cesium cathode

As a rule, K_g ranges between 5 and 10. The operating voltage for gas-filled phototubes is chosen so that no cumulative gas ionization (self-discharge) can occur in the tube. Otherwise, ions bombarding the cathode may destroy it.

For some time past, gas-filled phototubes have been finding fewer applications.

Response Curves. These show the relation between the sensitivity of a phototube and the illumination frequency. Figure 6-5 compares the response of a vacuum and a gas phototube. The sensitivity of the gas tube decreases with increasing frequency because ionization and recombination accompanying the gas discharge show a time lag owing to the low mobility of ions. The response characteristic of the vacuum phototube remains linear as far as 10^9 Hz. At the higher frequencies, the sensitivity falls because of the transit time of electrons.

Spectral Characteristics. The performance of a phototube depends on its spectral sensitivity, K_λ, that is, its response to a spectral (or monochromatic) light. This behaviour of phototubes, or, rather, of their cathodes, is represented by spectral characteristics. Such characteristics plotted for an antimony-cesium and an oxygen-cesium cathode appear in Fig. 6-6. The former is more sensitive to blue, and the latter to red light.

6-4. PHOTOMULTIPLIER TUBES

Definition. A *photomultiplier tube* (also called *multiplier phototube*) is an electron device in which photocurrent is multiplied by secondary emission. The construction of a photomultiplier tube is shown in Fig. 6-7.

When light strikes the photocathode, it emits electrons which, under the action of an electric field, move to the first dynode. It is

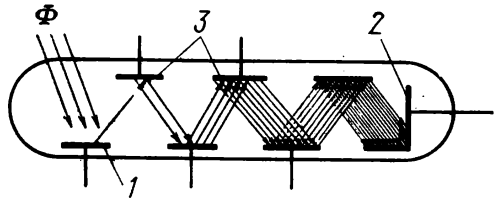


Fig. 6-7. Photomultiplier tube
1—photocathode; *2*—anode; *3*—secondary-emission cathodes (dynodes

coated with oxygen-cesium, copper-sulphur-cesium or any other composite material having the secondary-emission ratio $\sigma > 1$ (as a rule, σ ranges from 6 to 8).

As can be seen, the first dynode emits σ times the number of electrons ejected from the cathode, the second dynode emits σ^2 times the number of electrons ejected from the cathode, and so on. By the time the electron stream has reached the anode, the original photocurrent is amplified σ^n-fold, where n is the number of dynodes.

Current Gain of the Photomultiplier Tube. Strictly speaking, not all of the photoelectrons emitted by the photocathode, reach the

first dynode. The ratio of the number of electrons which reach the first dynode, n_1, to the number of electrons making up the photocathode current, n_{ph}

$$\eta_c = n_1/n_{ph} \tag{6-8}$$

is called the *collection efficiency*. Usually, η_c is slightly smaller than unity. Some electrons are also lost on their way between dynodes. This is expressed in terms of the *stage efficiency*, η, which is usually between 0.7 and 0.95.

Thus, if n_{ph} photoelectrons are emitted from the photocathode, the first dynode will collect $n_{ph}\eta_c$ electrons, the second dynode will collect $n_{ph}\eta_c\sigma_1\eta_1$ electrons, and so on. As a result, the anode will collect

$$n_a = n_{ph}\eta_c \prod_{i=1}^{n} \eta_i\sigma_i \tag{6-9}$$

electrons constituting the anode current $I_{ph,a}$.

The quantity

$$M = I_{ph,a}/I_{ph,k} \tag{6-10}$$

is called the *current gain of the photomultiplier tube.*

Integrated Sensitivity of the Photomultiplier Tube. The integrated sensitivity of the photomultiplier tube is equal to the product of the integrated sensitivity of the photocathode, K, and the amplification factor of the photomultiplier

$$K_{ph} = KM \tag{6-11}$$

and may be as high as several hundred amperes per lumen.

Construction. Photomultiplier tubes have found application in television, dosimetry, nuclear physics, astronomy, etc. Accordingly, they differ in construction.

Physically, a photomultiplier tube consists of two parts, namely an electron-optical structure arranged in an input chamber and a dynode structure.

The input chamber carries a photocathode formed by a semitransparent light-sensitive layer applied to the internal wall of the envelope, and a number of electrodes (the electron-optical structure) focusing the photoelectron stream towards the first dynode. Two likely designs of the input chamber are shown in Fig. 6-8. In the chamber of Fig. 6-8*a*, the photoelectron stream is focused by a cylinder cut away at an angle and attached to a diaphragm. In the other chamber (Fig. 6-8*b*), the stream is focused by electron lenses made up of cylindrical electrodes placed between the photocathode and the diaphragm.

Dynode structures are many and diverse. There exist dynode structures using electrostatic focusing, linear dynode structures and

structures with additional electrodes. A variant of the first structure type (using squirrel-cage dynodes) appears in Fig. 6-9*a*.

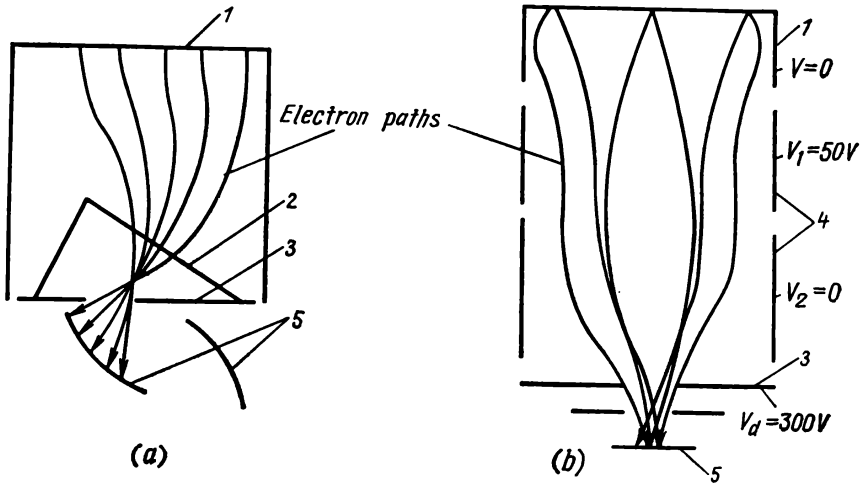


Fig. 6-8. Input chambers

(*a*) with a cylinder cut away at an angle; (*b*) with cylindrical focusing electrodes; *1*—photocathode; *2*—cylinder cut away at an angle; *3*—diaphragm; *4*—electrodes of the electron-optical structure; *5*—dynode

The linear dynode structure is shown in Fig. 6-9*b*. Here, the dynodes are arranged as a Venetian blind. Secondary electrons move to every next stage through the slots between the dynodes.

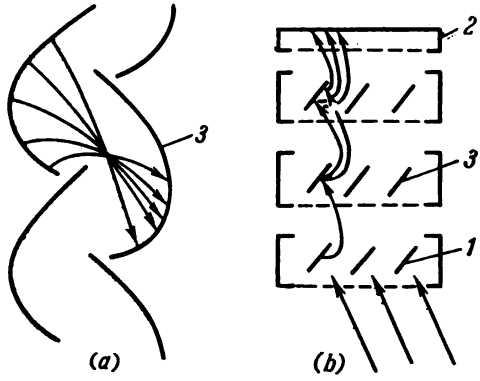


Fig. 6-9. Dynode structures

(*a*) with squirrel-cage dynodes; (*b*) linear; *1*—photocathode; *2*—anode; *3*—dynodes

Dynode structures of the third type have additional electrodes (for example, grids) designed and arranged so as to improve the focusing of the electron stream.

Chapter Seven

ELECTRON-BEAM DEVICES

7-1. CONSTRUCTION AND PRINCIPLE OF OPERATION OF CATHODE-RAY TUBES

Definition. *Electron-beam devices* are vacuum tubes utilizing an electron stream concentrated as a beam or beams. An electron-beam device made in the shape of a bulb stretched in the direction of the beam is known as the *cathode-ray tube* (abbreviated CRT).

The electron beam can be varied in position by electrostatic and magnetic fields, and in intensity, by electric fields. Cathode-ray tubes are used to produce a visible pattern from electric signals and also to store information carried by electric signals.

Electron-beam tubes used together with photocathodes may convert an optical image into a sequence of electric signals of a certain shape.

In an *electrostatic CRT* (Fig. 7-1*a*), the electron beam is focused and deflected by electrostatic fields. The CRT shown in Fig. 7-1*b* uses not only electrostatic, but also magnetic beam focusing and deflection. Accordingly, such tubes are known as *magnetic CRTs.*

In the Soviet Union, electron-beam devices are symbolized in accordance with a USSR State Standard, GOST 7624-62. The terminology and definitions are set up by GOST 17791-72.

Principle of Operation of the Electrostatic CRT. The envelope of an electrostatic CRT is a rather long, funnel-shaped bulb having in its neck a small cylindrical cathode enclosing a heater. The outside end of the cylinder carries an oxide layer; when heated, it emits electrons. The cathode is placed inside another cylinder having a small aperture in the bottom. This cylinder is termed the *modulating (control) grid*; it controls, or modulates, the electron beam emitted by the cathode. The modulating grid is held at a negative potential varied from zero to several tens of volts with respect to the cathode. Variations in this potential affect the density of the space charge near the cathode and, as a consequence, the height of the potential barrier. When the modulating grid is made more negative, a proportion of electrons return to the cathode and the density of the electron

stream is decreased. The electron stream is formed exclusively by the electrons passing through the diaphragm aperture 1 mm in diameter.

Next comes the anode structure which also consists of cylindrical electrodes, namely the focusing and accelerating anodes. The first anode held at a negative potential of a few hundred volts accelerates the electron stream emitted by the cathode. The second anode is at

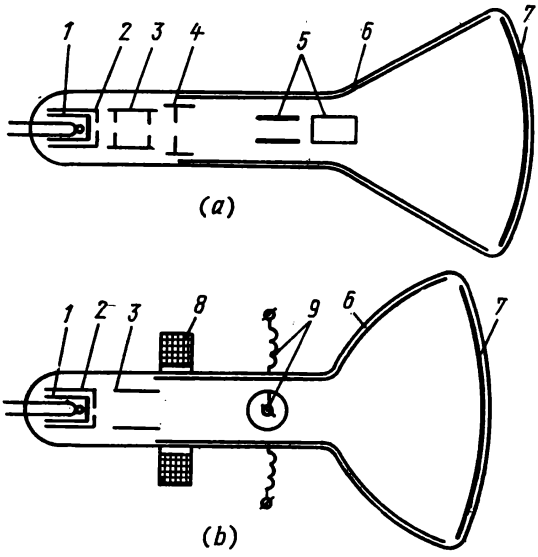


Fig. 7-1. Cathode-ray tube (*a*) electrostatic CRT; (*b*) magnetic CRT; *1*—cathode; *2*—modulating grid; *3*—first anode; *4*—second anode; *5*—deflection plates; *6*—conductive coating; *7*—screen; *8*—focusing coil; *9*—deflecting coils

a potential which, in some tubes, reaches tens of kilovolts, so the electrons leave the second anode at an appreciable velocity. Also, the anodes *focus* the beam. Since the cathode, modulating grid, and anodes are all held at different potentials, the electric fields between them are nonuniform in nature. Each field is said to form an *electron lens* (see Sec. 7-3). Electrons passing through the lenses are converged to form an electron beam at the screen. The whole electrode structure is mounted on supports and constitutes a single device known as the *electron gun.*

On leaving the electron gun, the electron beam founds itself in the *deflection structure* which, as its name implies, serves to deflect the beam in space. In electrostatic CRTs, it consists of two pairs of parallel plates set one forward of the other and at right angles to each other. An applied potential produces an electric field between each pair. By varying the applied potential, this field may be varied to cause the beam to deflect through a desired angle.

One pair, called the *horizontal-deflection plates*, establishes an electric field whose strength vector lies in a horizontal plane. Va-

riations in the potential difference between these plates deflect the beam horizontally. The other pair deflects the electron beam vertically, and these are the *vertical-deflection plates.*

On passing through the deflection plates, the electron beam travels in the funnel-shaped portion of the tube, and the electrons finally reach the screen. The screen is a thin layer of phosphor applied to the inner face of the tube. It fluoresces during bombardment by electrons. When the electron beam strikes the screen, it excites the phosphor and produces a small glowing spot visible through the tube glass. The better the electron beam is focused, the smaller the light spot and the narrower its trace on the screen. The brightness of the screen glow depends on the number and energy of electrons striking its surface per unit time. It can be varied by changing the voltage at the modulating grid or the second anode. The former alters the number of electrons, the latter affects their velocity.

Construction of the Magnetic CRT. A sketch of a magnetic cathode-ray tube is presented by Fig. 7-1*b*. Here, the electron gun consists of the same electrodes, except the second anode. The only electron lens is formed by the nonuniform electric field between the modulating grid and the first anode. The function of the second focusing lens is carried out by the nonuniform magnetic field established by a focusing coil put on the tube neck near the first anode. The beam is deflected by two pairs of deflection coils.

7-2. INTENSITY MODULATION OF THE BEAM

This is one of the basic processes in cathode-ray tubes. By varying the beam current density, we can control the brightness of the screen glow over a wide range from zero to a maximum specific for a given type of cathode-ray tubes.

The current density of the electron beam is controlled by varying the electric field in the electron gun between the cathode and the modulating grid (Fig. 7-2). As already noted, the modulating grid is a cylinder with a small aperture disc in the centre.

The modulating grid is held at a small negative potential, V_m, with respect to the cathode, which can be varied from zero to several tens of volts. Between the cathode and modulating grid, near the aperture disc, exists a nonuniform electrostatic field whose equipotential surfaces convex towards the cathode. Variations in the field shape markedly affect the current emitted by the cathode. This effect arises from variations in the space charge near the cathode and in the cathode surface area near which there is a field with a positive potential gradient. The modulating grid has a similar function as the control grid in the vacuum triode. When the modulating grid grows negative, the potential barrier near the cathode increases and the space charge builds up. Accordingly, one would think that the

cathode current depends on the modulating grid voltage as defined by the equation of the three-halves power law given in reference [12]

$$I_k = b\,(V_m - V_{m,\,c\text{-}o})^{3/2} \tag{7-1}$$

where V_m is the modulating grid voltage, $V_{m,c-o}$ is the cut-off voltage, that is, one at which the electron current is zero, and b is the constant of proportionality.

This, however, is not exactly the case. The cathode current undergoes an additional change due to variations in the cathode area threaded by the field with a positive potential gradient (accelerating field). If the modulating grid is at zero potential (Fig. 7-2*a*), almost all of the cathode surface is threaded by the positive potential gradient field. Figure 7-2*b* shows a map of the field when the potential at the modulating grid is close to the cut-off value. In this case, a positive gradient field exists only near an insignificant area at the cathode centre, the remaining surface being threaded by a negative gradient field. As a result, a decrease in the negative modulating grid voltage brings about a faster increase in cathode current than it follows from Eq. (7-1). The relation between the cathode current and modulating grid potential is defined by an equation given in reference [12]

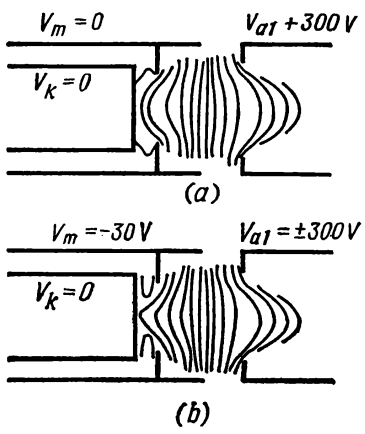


Fig. 7-2. Electric field near the modulating grid

$$I_k = b\,[(V_m - V_{m,\,c\text{-}o})/V_{m,\,c\text{-}o}]^{\gamma}\,|\,V_{m,\,c\text{-}o}\,|^{3/2} \tag{7-2}$$

Here, I_k is in microamperes, V_m and $V_{m,c-o}$ are in volts, b = (2.3 to 3) $\times$ 10^{-6} A/V$^{3/2}$, and γ ranges from 2.5 to 3.5.

In cathode-ray tubes, not all of the emitted electrons which have overcome the potential barrier near the cathode reach the screen. Some electrons are stopped by the limiting aperture discs of the first and second anodes because the beam markedly diverges as a result of focusing. Owing to this, the beam current past the focusing structure may be a fraction of I_k.

Modulation Characteristic. The relationship $I_k = f(V_m)$ is known as the *modulation characteristic* of a cathode-ray tube.

The modulation characteristics of the electrostatic and magnetic CRT's are shown in Fig. 7-3. As is seen, the relation between I_k and V_m is not the same.

In electrostatic CRTs, the modulation characteristic obeys the law given by Eq. (7-2) only near the cut-off potential when the beam

current is low. As the modulating grid becomes less negative, the emitting surface area of the cathode increases. As a result, some electrons are emitted at large angles, the beam spreads at the electrostatic lenses and the second anode, and a portion of electrons are intercepted by the limiting aperture disc. A further rise in the beam

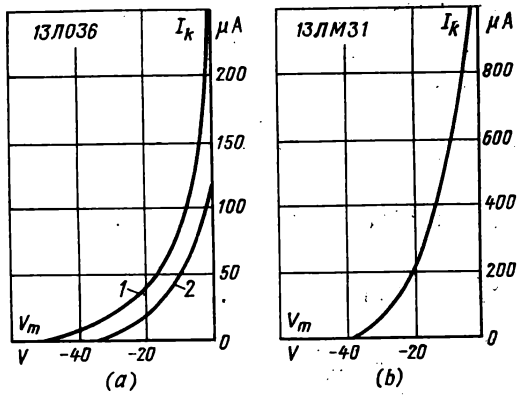


Fig. 7-3. Modulation characteristics of CRTs
(*a*) electrostatic CRT; *1*—at $V_{a_2} = 2.2$ kV; *2*—at $V_{a_2} = 1.5$ kV; (*b*) magnetic CRT

current may be obtained by increasing the density of the cathode current, that is, in accordance with the three-halves power law, Eq. (7-1).

Since the electron gun of a magnetic cathode-ray tube has no aperture discs, the beam current is close to the cathode current, and the relation between the cathode current and modulating grid voltage obeys Eq. (7-2).

The modulation characteristic depends on the cut-off voltage, $V_{m,c-o}$, and the nominal voltage, $V_{m,n}$, that is, the voltage at which the beam current reaches the value nominal for a given tube. The difference between these voltages

$$\Delta V_m = |V_{m,\,c-o}| - |V_{m,\,n}| \tag{7-3}$$

is called the *amount of modulation.*

The cut-off voltage depends not only on V_{a2}, but also on the voltages at the other electrodes. Present-day CRTs have an additional preaccelerating electrode placed between the modulating grid and the first anode (see Fig. 7-7*a*). As with vacuum tubes, the resultant field may be stated in terms of a composite controlling voltage

$$V_c = V_m + D_m V_{pe} + D_m D_{pe} V_a \tag{7-4}$$

where V_{pe} is the voltage at the preaccelerating electrode, V_a is the voltage at the first anode, and D_m and D_{pe} are the penetration fac-

tors of the modulating grid and the preaccelerating electrode, respectively.

In most electron guns, the effect of the third electrode is negligibly small, so we may write for the cut-off voltage

$$V_{m,\ c\text{-}o} = (\text{approx.}) - D_m V_{pe} \qquad (7\text{-}5)$$

The amount of modulation, ΔV_m, is related to the slope of the modulation characteristic measured in mA/V

$$S_m = \partial I_k / \partial V_m \qquad (7\text{-}6)$$

It is obvious that the amount of modulation decreases with increasing slope. As can be seen from Fig. 7-3, magnetic CRTs have a considerably higher S_m. Owing to this, magnetic cathode-ray tubes are used where a large amount of modulation is required, for example, in intensity-modulated radar indicators.

7-3. FOCUSING STRUCTURES. THE ELECTRON GUN

The electron gun has electrodes which not only control the intensity of the beam, but also focus the electron stream. So, before going into the details of typical electron guns used in CRTs of various purposes, we shall discuss the basic principles of focusing structures.

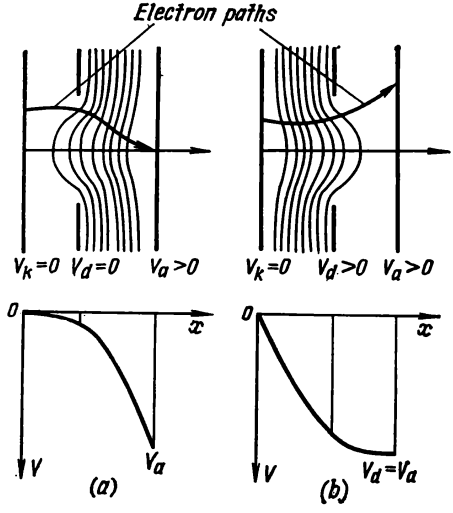


Fig. 7-4. Electron aperture lenses and potential distribution in them (*a*) converging lens; (*b*) diverging lens

Focusing of an electron stream into a narrow beam is based on the laws of electron motion in uniform and nonuniform electric and magnetic fields. These laws are dealt with in a course on electron optics which widely uses the concepts of light optics.

Electron Lenses. This term applies to nonuniform axial-symmetric electric fields employed in cathode-ray tubes to control the velocity and direction of electrons. Electron lenses may also be composed by axial-symmetric uniform and nonuniform magnetic fields. In electron optics we have *aperture lenses, unipotential lenses, immersion projective lenses* and *immersion objective lenses.*

Aperture lenses are shown in Fig. 7-4. The figure also illustrates the potential distribution along the lens axis. Aperture lenses may be either converging or diverging. In the aperture plane of the converging lens, $\partial^2 V_d / \partial x^2 > 0$; in the diverging lens, $\partial^2 V_d / \partial x^2 < 0$.

A unipotential lens is formed by a group of aperture lenses, the outer ones being at a constant and equal potential. A unipotential lens is shown in Fig. 7-5*a*.

The immersion projective lenses most commonly used in CRTs have cylindrical electrodes of the same or different diameters. The potentials at both ends of a lens are constant but different in magnitude. Immersion projective lenses appear in Fig. 7-5*b* and *c*.

An immersion objective lens is a structure consisting of a cathode whose image is formed, and aperture discs or cylinders which establish an accelerating field near the cathode.

As is known from light optics, the linear dimensions of an image produced by a spherical surface dividing two media are defined by the Lagrange-Helmholtz equation

$$y_2 = y_1 (\theta_1 n_1)/(\theta_2 n_2) \qquad (7\text{-}7)$$

where y_1 and y_2 are the linear dimensions of the object and image, respectively, θ_1 and θ_2 are the aperture angles, and n_1 and n_2 are the refractive indexes of the two media.

For an electron-optical system, the interface between two media is an equipotential surface which approximates a spherical surface in practical systems.

As in light optics, the refractive index in electron optics is the ratio of the electron velocity to the velocity of light. So in Eq. (7-7), instead of n_1 and n_2 we should substitute the respective electron velocities in the two media or their equivalent potentials

$$y_2 = y_1 \theta_1 \sqrt{V_1}/\theta_2 \sqrt{V_2} \qquad (7\text{-}8)$$

The size of an image can be reduced by increasing the ratio θ_2/θ_1, reducing the size of the object, y_1, and the relative velocity of the electrons ahead of and past the optical structure. If an optical structure were to use only one lens, these requirements would be rather difficult to meet. The point is that the object is the emitting surface of the cathode. Any reduction in its size would inevitably bring about a decrease in the beam current, and the size of the image (the light spot on the screen) would depend on the modulating grid potential. The angle, θ_1, at which electrons leave the cathode depends on the thermal velocities of the electrons, so it cannot be reduced significantly. Therefore, CRTs use focusing structures consisting of two or three lenses where the object for the second lens is the *crossover*, or the point past the first lens where the cross-sectional area of the beam is minimal.

The first, or condenser, lens should be electrostatic because it not only forms the crossover, but also accelerates the electrons emitted by the cathode. The second lens in a two-lens electron gun, called the *projective* or *projector lens*, is used to form an image of the crossover on the screen. In a three-lens electron gun, the projector lens is a third lens. In contrast to the condenser lens, the projector lens may

be electrostatic or magnetic. In the former case, the focusing structure is called electrostatic; in the latter case, the electron gun is said to have a magnetic focusing structure. In either case, the projector lenses should meet the same requirements: they must produce the

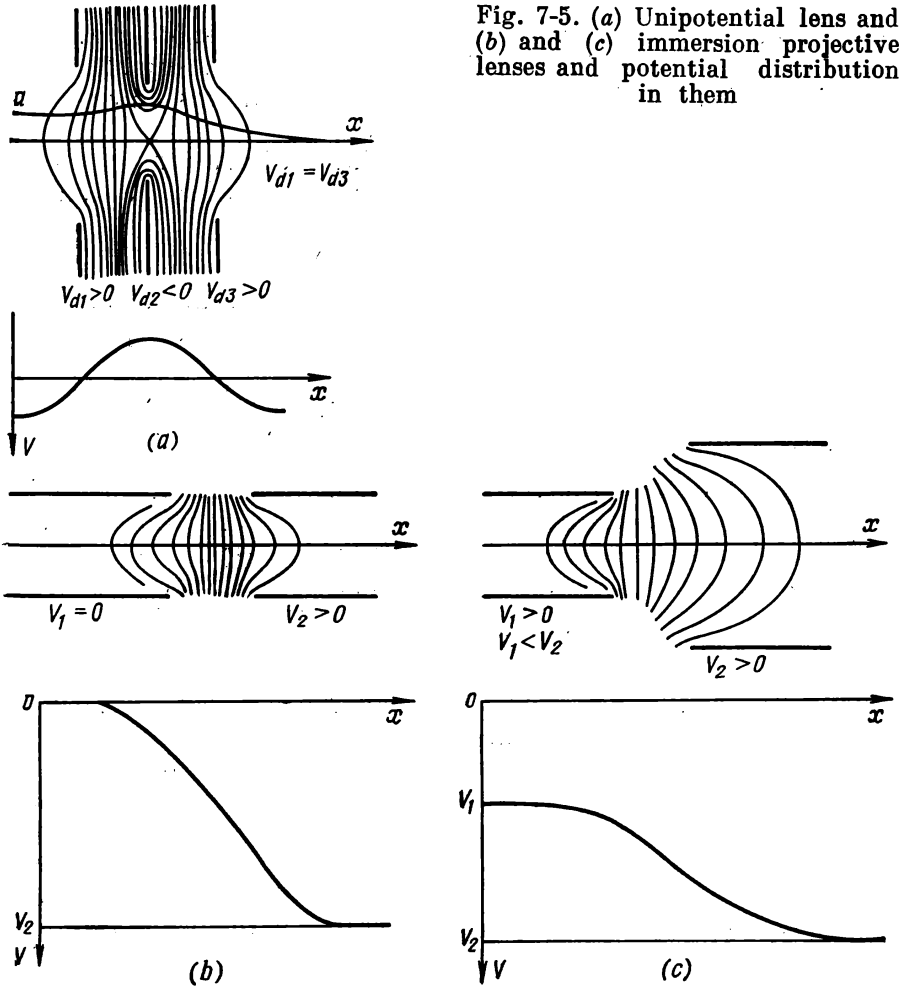


Fig. 7-5. (*a*) Unipotential lens and (*b*) and (*c*) immersion projective lenses and potential distribution in them

narrowest possible electron beam and, as a consequence, the smallest possible light spot on the screen. For this purpose, when designing electron guns, one seeks to reduce the size of the crossover past the condenser lens and the aperture angle θ_1' on the object side, increase the aperture angle θ_2 on the image side, and also reduce the electron velocity near the object and increase it near the screen.

Present-day focusing structures produce light spots less than 0.1 mm in diameter.

Electron Gun Using Electrostatic Focusing. Figure 7-6*a* shows an electron gun having three electrodes (a triode electron gun), namely a modulating grid, a first and a second anode. Its electrostatic focusing structure consists of an immersion objective lens (made up of the cathode, modulating grid and first anode) and an immersion projective lens (composed by the first and second anodes). The equi-

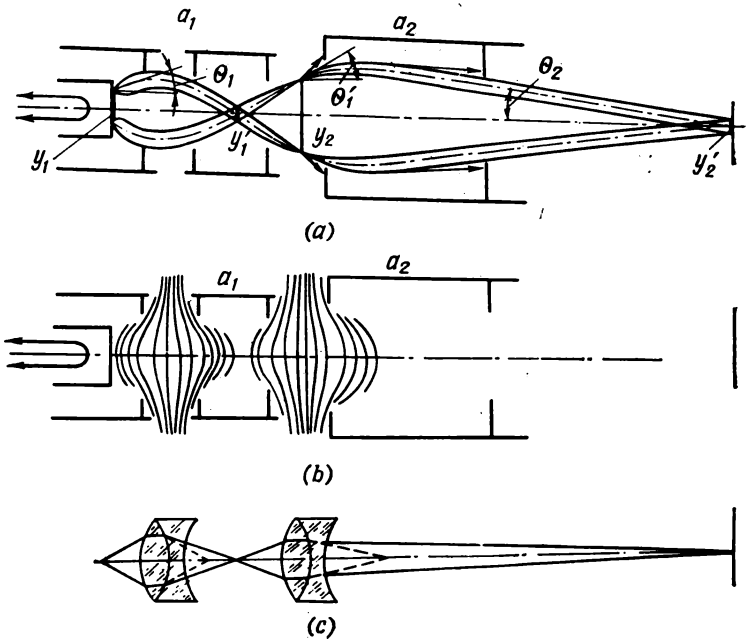


Fig. 7-6. Triode electron gun
(*a*) electron paths; (*b*) electric field; (*c*) light-optic analogy of electron lenses

potential lines of the field established by the two lenses of the structure are shown in Fig. 7-6*b*, their light-optical analogy appears in Fig. 7-6*c*. The first lens is composed by the nonuniform field between the modulating grid and the first anode, the second by the field between the first and second anodes. Each may be regarded as a combination of two light-optical lenses, namely a biconvex converging lens made up of the equipotential surfaces convexing to the cathode, and a biconcave diverging lens formed by the equipotential surfaces concaving in the opposite direction. The fields are shaped in such a way that the converging lens is more powerful than the diverging one. This is also promoted by the fact that, since the potential in the field of the diverging lens increases, the electron velocities in this field exceed those in the field of the converging lens. Thus, the resultant action of each of these composite lenses is to converge the beam; the diverging section of each lens helps only to increase the focal distance of the whole structure. In order to make the second

crossover form in the plane of the screen, the focal distance of the second lens may be varied by changing the refracting force of one of the lenses. This can be achieved by varying the voltage at either the first or the second anode. The first anode, however, is usually held at a much lower potential than the second; also, the circuit of the second anode handles heavy currents, so the second anode voltage is not used to control the beam focusing.

A major drawback of the triode electron gun is that the current (brightness) control inevitably affects beam focusing and vice versa.

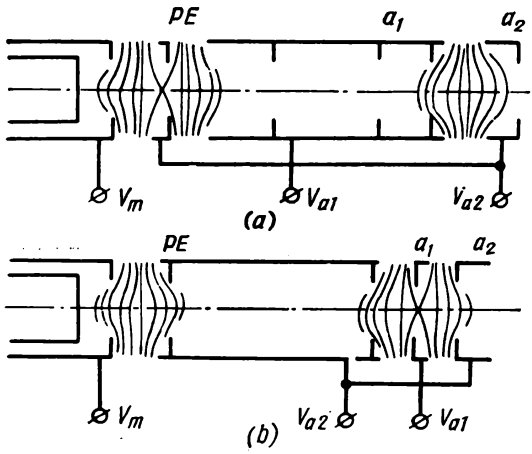


Fig. 7-7. Electron guns
(*a*) with a preaccelerating electrode; (*b*) with zero first-anode current

As the first anode voltage is varied to focuse the beam, the field near the cathode and, as a consequence, the beam current are also changed. On the other hand, as the modulating grid voltage is varied to control the brightness, the beam focusing is upset. To avoid this, present-day cathode-ray tubes utilize more sophisticated electron guns.

Figure 7-7*a* shows an electron gun which has a preaccelerating electrode placed between the modulating grid and the first anode and held at the second anode potential. Thus, this electron gun consists of an immersion objective lens and a unipotential lens. This not only reduces the interaction between the brightness control and beam focusing, but also improves the beam focusing because the beam radius at the crossover decreases. Figure 7-7*b* shows an electron gun with a zero-current first anode. As with the electron gun shown in Fig. 7-7*a*, the first anode whose function is to focuse the beam is separated from the modulating grid by a preaccelerating electrode. This electrode is held at a constant potential and serves as an electric screen between the first and second lenses, thereby preventing them from affecting each other. An important advantage of this elec-

tron gun is that no electrons fall onto the first anode which is made in the form of a large-diameter aperture disc, and the current in its circuit is close to zero. Thus, variations in the first anode potential have no effect on the current from the rectifier supplying all tube electrodes. Even with a low-power rectifier there is no interaction between the potentials at the various electrodes.

In these electron guns, the first anode is usually maintained at a potential of a few hundred volts, while the second anode and the preaccelerating electrode are held at several kilovolts. The reduction in the first anode potential makes it possible to compress the light spot on the screen, as follows from Eq. (7-8).

Figure 7-8 shows a *tetrode electron gun*. It has an immersion objective lens (made up of the cathode, the modulating grid and the first preaccelerating electrode), a low-power intermediate projector lens (formed by the first and second preaccelerating electrodes), and a main projection (unipotential) lens formed by the field between the second preaccelerating electrode and the second anode.

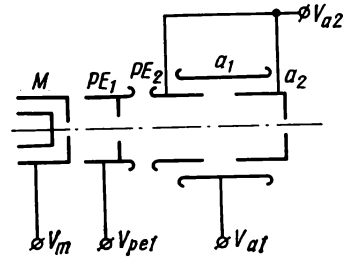


Fig. 7-8. Tetrode electron gun

The first preaccelerating electrode is held at a relatively low potential (about several hundred volts) which practically determines the cut-off voltage. This is also aided by the construction of the aperture disc generally used in the first preaccelerating electrode. The aperture disc reduces the effect that the potentials at the succeeding electrodes may have on the field near the cathode. Thus, the cut-off voltage in the tetrode electron gun is reduced considerably. Another advantage of the tetrode electron gun is a lower angular divergence of the beam past the crossover.

The choice of a particular type of electron gun or electrode structure depends on the purpose of a CRT and requirements imposed upon the quality of focusing, brightness, maximum distortion, etc.

Electron Gun Using a Magnetic Lens. The second lens in magnetic CRTs is established by the nonuniform magnetic field developed by a short coil, that is, one whose length is comparable with its inner diameter. Electron guns using magnetic lenses are shown in Fig. 7-9.

As in an electrostatic electron gun, the immersion objective lens is formed by the electric field of the electrodes. In a triode electron gun, it is made up of the cathode, the modulating grid and the anode; in a tetrode gun, it is composed of the cathode, the modulating grid and the preaccelerating electrode. The tetrode type also has a low-power immersion (intermediate projector) lens established by the field between the preaccelerating electrode and the anode. In both cases, the projector lens is magnetic.

Electron paths in the electrostatic lenses are the same as in an electrostatic gun (Fig. 7-6*a*). The crossover serves as an object for the magnetic projector lens formed by the nonuniform field of the short coil put on the neck of the tube.

Figure 7-10 illustrates how the nonuniform field of the short coil focuses the electron beam. In the general case, the electron velocity vector is at an angle α to the coil axis. Let us resolve the magnetic

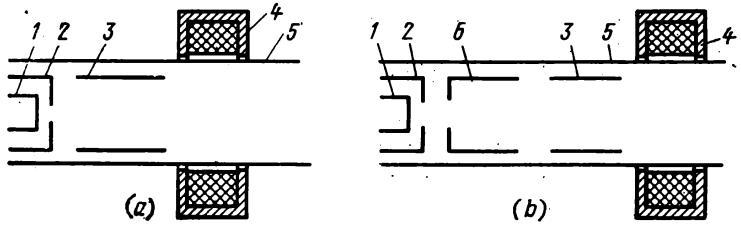


Fig. 7-9. Electron guns with a magnetic lens
(*a*) triode electron gun; (*b*) tetrode electron gun; *1*—cathode; *2*—modulating grid; *3*—anode; *4*—focusing coil; *5*—envelope; *6*—preaccelerating electrode

field induction vector, B, at point A into the radial, B_r, and axial, B_a, components. Since the angle α is small, and cos α is unity very

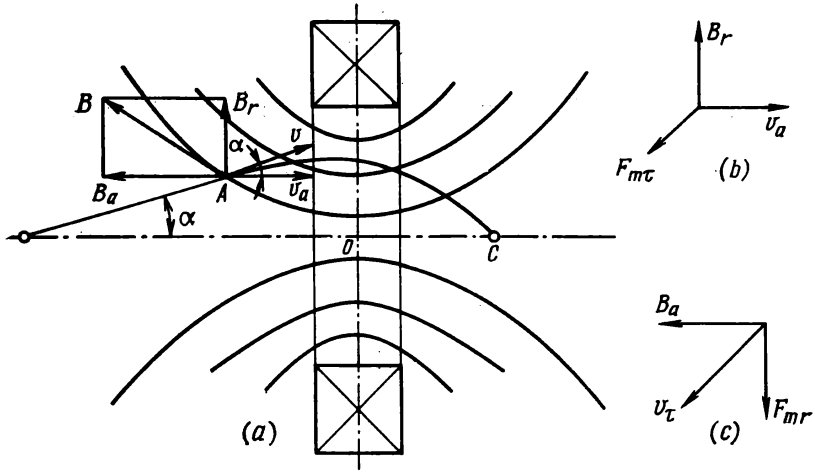


Fig. 7-10. Magnetic lens formed by a short-coil field

nearly, we may take it that $v_a = v \cos \alpha =$ (approx.) v. The electron is acted upon by a force $F_{m\tau}$ which is at right angles to the plane of the drawing. This force makes the electron revolve around the coil axis thereby giving rise to the tangential component, v_τ, of the electron velocity (Fig. 7-10*c*). This component, together with B_a, constitutes the force, F_{mr} which drives the electron to the axis. Thus, in a nonuniform magnetic field, the electron follows a helix

having a decreasing radius, which converges with the coil axis at a certain point C. Adjusting the current in the focusing point and, as a consequence, changing the magnetic induction, we may place this point of convergence at the screen, that is, focus the electron beam.

To improve the focusing action of the magnetic lens and also to concentrate the magnetic field in a limited space, the focusing coils are enclosed in magnetic screens. These are known as *screened coils*.

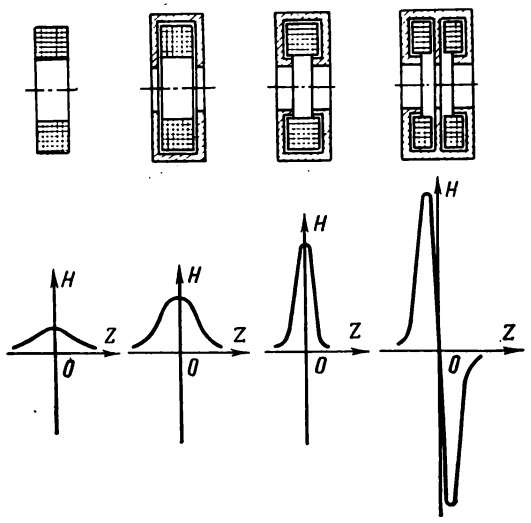


Fig. 7-11. Different types of magnetic focusing coils and their fields

Screens of different constructions are shown in Fig. 7-11 which also illustrates the curves of the magnetic field strength distribution along the tube axis.

Magnetic focusing structures offer a better focusing and lower image distortion. Their disadvantage is that they require powerful supply sources for the focusing coils, because the necessary magnetic fields can be developed only by high currents. Magnetic focusing has found wide application in radar CRTs.

7-4. DEFLECTION STRUCTURES

Electrostatic Deflection. The simplest electrostatic deflection structure consists of two pairs of parallel plates set one forward of the other (Fig. 7-12*a*). One pair of plates deflects the electron beam horizontally, the other, vertically. If we neglect the edge effect, the field between the plates may be deemed uniform and of strength $\mathscr{E} = V_p/d$.

Let us examine an electron travelling under the action of the field $\mathscr{E}$. The velocity of an electron approaching the deflection plates

depends on the voltage at the second anode

$$v_0 = \sqrt{2\,(e/m)\,V_{a2}} \tag{7-9}$$

At the entrance edges of the plates, its velocity along the z-axis is zero. Between the plates, the force $F = e\mathscr{E}$ makes the electron move

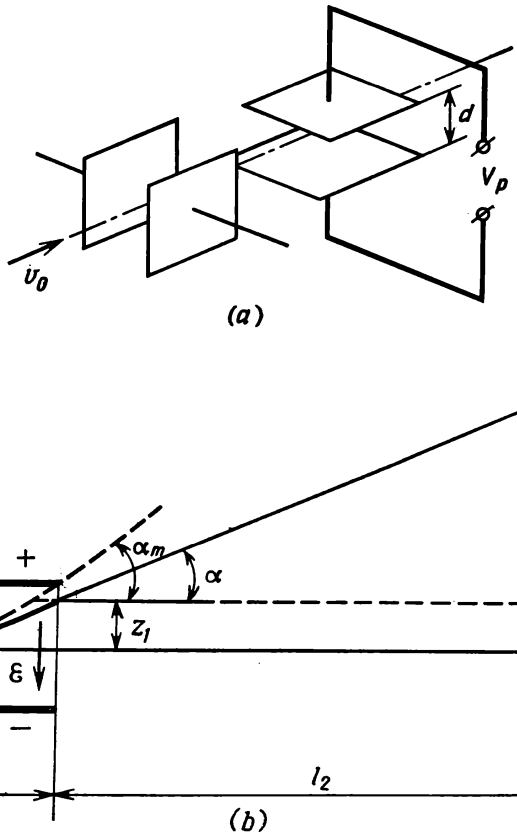


Fig. 7-12. (*a*) Deflection plates and (*b*) electron path

along the z-axis at a uniform acceleration. Consequently, its motion between the plates can be described thus

$$\left.\begin{aligned} x &= v_0 t \\ z &= (e\mathscr{E}/2m)\,t^2 \end{aligned}\right\} \tag{7-10}$$

Finding the time from the first equation, $t = x/v_0$, and substituting it into the second equation, we get

$$z = (e\mathscr{E}/2mv_0^2)\,x^2 \tag{7-11}$$

By the time it leaves the plates, the electron moving between them along a parabola will have been deflected through a certain distance

z_1. On leaving the plates, it will move towards the screen along a tangent to its curved path. The deflection of the electron during the time it travels from the plates to the screen is z_2. The total deflection from the tube axis is $z = z_1 + z_2$. The deflection z_1 can easily be deduced from the equation of the parabolic electron path, Eq. (7-11)

$$z_1 = (e\mathscr{E}/2mv_0^2)\, l_1^2 \tag{7-12}$$

Substituting Eq. (7-9) and $\mathscr{E} = V_p/d$ into this equation, we get

$$z_1 = V_p l_1^2/4dV_{a2} \tag{7-13}$$

Deflection z_2 is equal to $l_2 \tan \alpha$. But $\tan \alpha = dz/dx$ is the slope of the electron path at the point where it leaves the plates

$$\tan \alpha = dz/dx = \frac{d}{dx}\left(\frac{e\mathscr{E}}{2mv_0^2} x^2\right) = (e\mathscr{E}/mv_0^2)\, x \tag{7-14}$$

Substituting the expressions for $\mathscr{E}$ and v_0 into this equation, we get

$$\tan \alpha = (V_p/2dV_{a2})\, l_1 \tag{7-15}$$

Thus,

$$z_2 = V_p l_1 l_2/2dV_{a2} \tag{7-16}$$

The total deflection of the electron from the tube axis is

$$z = z_1 + z_2 = (V_p l_1/2dV_{a2})\,(l_1/2 + l_2) \tag{7-17}$$

The sum in the brackets is the distance, L, from the centre of the plates to the screen. So

$$z = V_p l_1 L/2dV_{a2} \tag{7-18}$$

Deflection Sensitivity. Equation (7-18) shows that the deflection of the electron beam is directly proportional to the potential difference between the deflection plates. The proportionality factor of this relation which gives the displacement of the beam on the screen per volt of deflection potential is called the *deflection sensitivity* of a CRT

$$\varepsilon_e = l_1 L/2dV_{a2} \tag{7-19}$$

The deflection sensitivity depends on the size of the deflection plates; it is inversely proportional to the squared electron velocity and directly proportional to the distance between the plates and the screen. As a rule, the horizontal and vertical deflections differ because of different distances between the vertical and horizontal plates and the screen.

In the beginning of our discussion, we neglected the edge effect and assumed that the field between the plates was uniform. Actually, however, the field is bent at the plate edges and affects an approaching electron already when it is at some distance from the plates. Therefore, in practice the deflection sensitivity is 15% above the computed value.

In addition to the deflection sensitivity which is one of the most important parameters of a CRT, the deflection structure is characterized by the maximum deflection angle, α_m.

As can be seen from Fig. 7-12*b*, with parallel deflection plates, α_m is limited by the length of, and the spacing between, the plates. To increase α_m, cathode-ray tubes use diverging, bent or curved deflection plates (Fig. 7-13). The best results are achieved with curved plates. Here, the electron path is equidistant from the surface of the deflection plate.

The deflection plates shaped as shown in Fig. 7-13 also increase the deflection sensitivity. If we take the deflection sensitivity of a

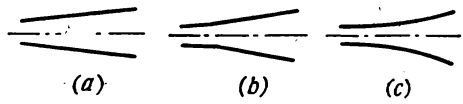


Fig.7-13. Different shapes of deflection plates
(*a*) diverging (*b*) bent (*c*) curved

CRT using curved deflection plates as unity, then all other things being equal, the deflection sensitivities of CRTs using bent, diverg-

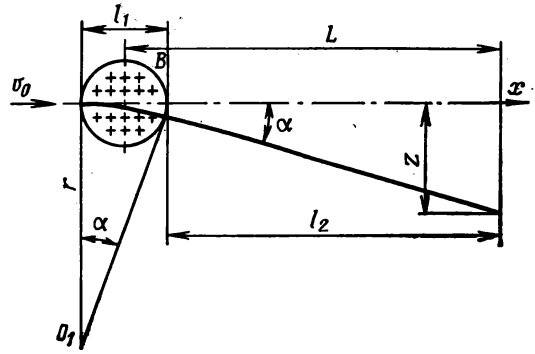


Fig. 7-14. Magnetic deflection of the electron beam

ing and parallel deflection plates will be 0.95, 0.84 and 0.51, respectively.

Magnetic Deflection. Magnetic deflection of the beam is effected by two pairs of deflecting coils placed around the neck of a CRT and developing magnetic fields at right angles to each other. Let us examine how an electron is deflected by the magnetic field of one pair of coils, assuming that the field is limited by the coil diameter and is uniform within this space. In Fig. 7-14, the lines of force of the magnetic field are shown receding from the reader at right angles to the plane of the drawing. As is known from a course on physics,

an electron travelling with the initial velocity v_0 in a magnetic field whose induction vector B is normal to the velocity vector v_0 follows a circle whose radius is

$$r = mv_0/eB \tag{7-20}$$

On leaving the magnetic field, the electron continues moving along the tangent to its curved path at the point where it leaves the field. The electron deflects from the tube axis through $z = L \tan \alpha$. At small angles, α = (approx.) $\tan \alpha$, so z = (approx.) $L\alpha$.

The central angle is $\alpha = s/r$ = (approx.) l_1/r, where s is the curve along which the electron moves in the field B. Substituting the expression for r into this equation, we get

$$\alpha = (eB/mv_0)\, l_1 \tag{7-21}$$

Thus, the deflection of the electron is

$$z = (eB/mv_0)\, Ll_1 \tag{7-22}$$

Expressing the electron velocity, v_0, in terms of the anode voltage, we obtain

$$z = \sqrt{e/2m}\,(Ll_1/\sqrt{V_a})\,B \tag{7-23}$$

Recalling that the magnetic induction is proportional to the ampere-turns, wI, we may write

$$z = \sqrt{e/2m}\,(Ll_1/\sqrt{V_a})\,kwI \tag{7-24}$$

Deflection Sensitivity of a Magnetic CRT. The proportionality factor connecting the beam deflection and the magnetic induction is termed the *deflection sensitivity of a magnetic CRT*

$$\varepsilon_m = \sqrt{e/2m}\, Ll_1/\sqrt{V_a} \tag{7-25}$$

The deflection sensitivity and, what is more important, its constancy for any deflection angles substantially depend on the deflecting-coil design.

Deflecting-Coil Design. Iron-cored deflecting coils (Fig. 7-15*a*) make it possible to concentrate the magnetic density within the required space. Iron-cored coils are used only at low-frequency deflection voltages, because, as the frequency of deflection voltage rises, core losses increase, too. Television and radar CRTs generally employ air-cored deflecting coils. Such coils are shown in Fig. 7-15*b* and *c*. In order to make the magnetic field as uniform as possible, the coil edges are flared and the coils themselves are given the shape of the tube neck. The coils are arranged in a nonuniform manner, the number of turns at the edges being usually two or three times that at the centre. To reduce the leakage magnetic flux, air-cored coils are enclosed in steel screens.

Advantages and Disadvantages of Electrostatic and Magnetic Deflection. As is seen from Eq. (7-23), magnetic deflection depends on the electron velocity to a lesser extent than electrostatic deflection, Eq. (7-18). Because of this, magnetic deflection has found application in CRTs with high anode potentials necessary to produce very bright light spots.

A limitation of magnetic deflection is that it cannot be utilized at deflection voltages with frequencies higher than 10 to 20 kHz,

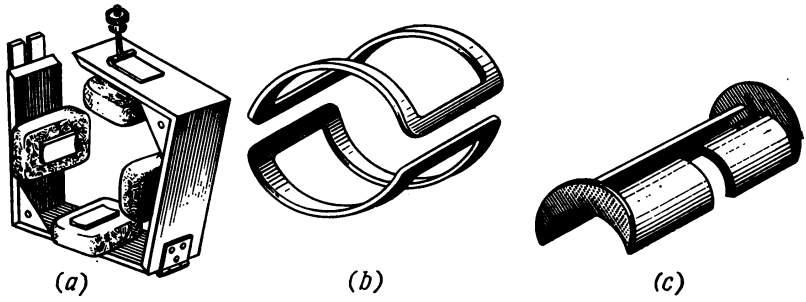


Fig. 7-15. Deflecting coils
(*a*) with an iron yoke; (*b*) and (*c*) without a yoke

whereas conventional electrostatic tubes can be used at frequencies as high as tens of megahertz or even higher. Besides, magnetic deflecting coils dissipate much power and, as a consequence, require powerful supply sources.

An advantage of magnetic deflection is that the deflecting coils are placed outside the tube, so the deflection structure (or the deflection yoke) can be rotated round the tube axis.

7-5. CRT SCREENS

On passing through the deflection structure, electrons strike the screen which is a layer of a phosphor applied to the inner face of the cathode-ray tube by deposition from a suspension or by evaporation.

Phosphors. These are materials capable of emitting light when struck by electrons.

Most of the phosphors used in commercial tubes are oxides and sulphides of zinc, silicon, cadmium, magnesium, etc. These materials are mixed with silver, manganese or chromium which form local energy levels—recombination centres—in the forbidden band of the stock material. The electrons of the phosphor absorb the energy of the emitted electrons and move from the impurity and valence bands to the conduction band. Free electrons occupy this high-energy level for a short time. As they interact with the crystal lattice, they lose their energy and gradually return to the local impurity levels

and the valence band. Under certain conditions such transitions are accompanied by emission of light quanta, or photons. This is known as *cathodoluminescence*. Absorption and emission of light by semiconductors are discussed in more detail in Chapter 14.

The phosphors of CRTs are expected to satisfy some special requirements. They should have high luminous efficiency, that is, they should effectively convert the kinetic energy of electrons into light. In addition, the phosphors should be properly degassable, stand up well to temperature and have a long service life. Also important are the colour of fluorescence and the persistence, that is the length of time that a phosphor glows after the exitation is removed.

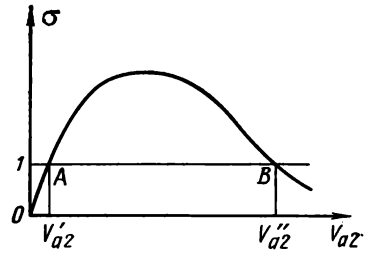


Fig. 7-16. Secondary-emission ratio of a phosphor as a function of second-anode voltage

Screen Potential. When the screen is bombarded by electrons, phenomena characteristic of secondary emission take place (see Sec. 1-3), so the screen emits secondary electrons. These electrons should be removed from the space near the screen surface, since otherwise, they would return to the screen, form a negative space charge, and apply a negative potential to the screen. As a result, the primary electrons near the screen surface would be slowed down and the screen would cease to glow. To remove secondary electrons, an electric field should be developed near the screen. For this purpose, the tube walls near the screen are given a coat of a current-conducting material which is connected to the second anode inside the tube and allows secondary electrons to leak off the screen.

The screen potential depends on the secondary-emission ratio. Referring to the curve relating the secondary-emission ratio to the accelerating voltage (Fig. 7-16), at low acceleration potentials (portion OA), σ is smaller than unity. In this case, electrons accumulate at the screen and its potential at any acceleration voltage from 0 to V'_{a2} tends to reach the cathode potential which is taken to be zero. As a result, the screen ceases to glow. This value, V'_{a2}, is called the *first critical potential.*

As the accelerating voltage grows (portion AB), the secondary-emission ratio exceeds unity, and the screen becomes positive with respect to the second anode. But the screen potential is only slightly higher than the anode potential, because some secondary electrons return to the positive screen thereby reducing its potential. When the screen potential is approximately equal to that at the second anode, a dynamic equilibrium takes place, that is, the number of secondary electrons leaving the screen is equal to the number of primary electrons in the beam.

At $V_{a2} > V''_{a2}$, the secondary-emission ratio is below unity, and, as a consequence, the number of electrons leaving the screen is smaller than that of primary electrons. The accumulation of negative charge brings the screen potential down to V''_{a2}, at which $\sigma = 1$. Thus, V''_{a2} corresponds to a dynamic equilibrium when the number of electrons leaving the screen equals the number of primary electrons. The value of V''_{a2} is called the *second critical* or *limiting potential* of the phosphor.

Brightness of a Phosphor. The brightness of a phosphor, B, obeys the Lenard law

$$B = Aj(V - V_0)^n \tag{7-26}$$

and depends on the electron-beam density, j, the potential difference between the cathode and the screen, V, and a minimum potential, V_0, at which the phosphor begins to fluoresce. The coefficient A and the exponent n which ranges from 1 to 3 depend on the properties of the phosphor. The value of V_0 varies from phosphor to phosphor and may be anywhere between 10 and several hundred volts.

Luminous Efficiency. As already noted, the screen glows because the electrons of the phosphor absorb the energy of primary electrons. However, far from all of the kinetic energy of primary electrons is converted into the energy of visible radiation. A considerable part of it is expended to heat the screen, knock out secondary electrons and produce radiation in the invisible portion of the spectrum. The efficiency of a phosphor, defined as the ratio of the energy of visible radiation to the power expended to excite it is called the *luminous efficiency* of the phosphor. The energy of the visible radiation is usually expressed in terms of the luminous flux or luminous intensity, so the luminous intensity is measured in candelas per watt. For CRT phosphors, the luminous efficiency ranges between 0.1 and 10 cd/W, and increases with increasing kinetic energy of electrons. That is why, to compare various phosphors, the luminous efficiency is measured at a specified accelerating voltage.

Experience shows that the luminous efficiency of a phosphor gradually decreases under cathode-ray bombardment. In some cases, a short period of "rest" may be enough for the phosphor to recover. After a long time in use, however, this deterioration in efficiency, known as *electron burn*, becomes irreversible.

Persistence. The kinetic energy of the impinging electrons takes some time, usually a fraction of a microsecond, to turn into the energy of visible radiation. Likewise, after the electron beam has been turned off, the phosphorescence takes some time to die away. It decays exponentially, with the exponent being a function of the physical and mechanical properties of the phosphor. The time interval between the instant when the electron beam is turned off and the instant when the intensity of phosphorescence has fallen to one per

cent of its initial brightness is called the *persistence of the phosphor*. Phosphors may be classed into very short persistence (less than 10^{-5} s), short persistence (10^{-5} to 10^{-2} s), medium persistence (10^{-2} to 10^{-1} s), long persistence (10^{-1} to 16 s) and very long persistence types (over 16 s). Short and very short persistence types are used in oscilloscopic CRTs, and those with medium persistence are used in

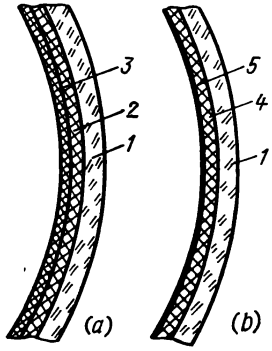


Fig. 7-17. Composite screens
(*a*) double-layer (long persistence screen); (*b*) aluminized screen; *1*—glass; *2*—phosphor with yellow phosphorescence and long persistence; *3*—phosphor with blue phosphorescence and short persistence; *4*—phosphor; *5*—aluminium

television CRTs. Radar indicator tubes usually employ long-persistence phosphors.

Composite Screens. Long persistence is obtained by using *double-layer screens* (Fig. 7-17*a*). At first, the glass is given a coat of a phosphor with yellow phosphorescence and long persistence. Then, it is covered with a layer of a phosphor having blue phosphorescence and short persistence. Electrons bombarding the screen excite the atoms of the first layer and produce blue fluorescence. The first layer excites the second layer, and the latter begins to fluoresce, too. With such an excitation, the persistence of the second layer depends, in particular, on the thickness of the first layer and is generally much longer than when it is excited by cathode-ray bombardment.

Nowadays, use is widely made of *aluminized* (or *aluminium-backed*) *screens*. In these, the phosphor layers are given a coat of aluminium about one micrometre thick (Fig. 7-17*b*). As a rule, the aluminium coating is connected to the second anode, so the screen potential is always equal to V_{a2} and does not depend on the secondary-emission ratio. It is most advisable to use aluminized screens in tubes operating at high voltages, because an increase in the accelerating voltage and, as a consequence, in screen potential raises the luminous efficiency of the screen and reduces the absorption of beam electrons by the aluminium layer. Also, the aluminium backing raises the luminous efficiency by reflecting the luminous flux towards the observer, and also improves the picture contrast by reducing mirror reflection of light from the funnel-shaped portion of the envelope and interference between neighbouring parts of the screen. Of fundamental

importance is the colour of phosphor fluorescence. Oscilloscopic CRTs use phosphors having green fluorescence which is least tiresome for the human eye. Such fluorescence is produced by zinc silicate (willemite) activated with manganese. The spectral characteristic of willemite and the response curve of the human eye are shown in Fig. 7-18*a*. When it is necessary to photograph a picture, the best

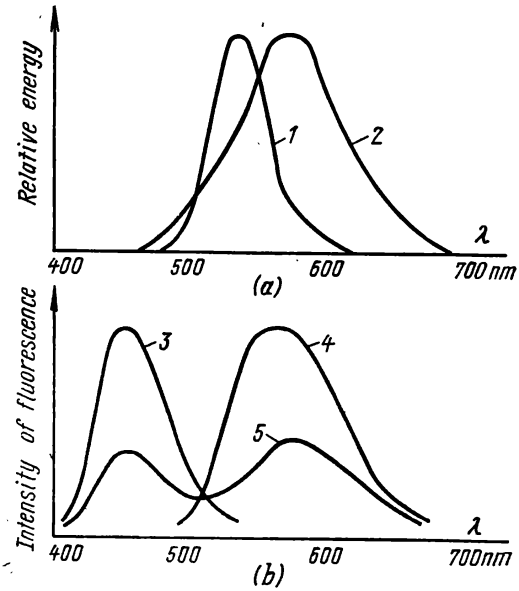


Fig. 7-18. (*a*) Spectral characteristic of a phosphor with green fluorescence and (*b*) spectral characteristic of a phosphor with white fluorescence

1—spectral characteristic of willemite; *2*—human eye response curve; *3*—spectral characteristic of the blue component; *4*—spectral characteristic of the yellow component; *5*—characteristic of the phosphor with white fluorescence

results are achieved with blue-fluorescence phosphors such as calcium tungstate. Television picture tubes employ phosphors with white fluorescence, for this purpose, they use a combination of two phosphors having blue and green fluorescence. The spectral characteristics of a television picture tube and its components are shown in Fig. 7-18*b*.

7-6. CONSTRUCTION AND TYPES OF CATHODE-RAY TUBES

Cathode-ray tubes are widely used to observe electric signals visually (in oscilloscopes or radar displays), transmit and receive images by means of television, record and store electric signals in memory devices, etc. The parameters and construction of CRTs vary in accordance with their application.

Oscilloscopic CRTs. These are intended to produce the patterns of electrical processes varying with time. For this purpose, use is commonly made of electrostatic CRTs. Electrostatic deflection does not require power from the source of the signal to be viewed, which is very important when we study signals generated by low-power sources. This also makes it possible to apply higher frequency deflection voltages.

Any electrical process may be displayed on the CRT screen by producing a pattern of, say, voltage as a function of time. To do this, a linearly rising voltage having a sawtooth shape and called the *sweep voltage*, v_s (Fig. 7-19*a*), is applied to the horizontal deflection plates which are closer to the screen and have a lower sensitivity.

When v_s is zero, the electron beam is at the leftmost edge of the screen. Driven by the rising sweep voltage, the beam traces out

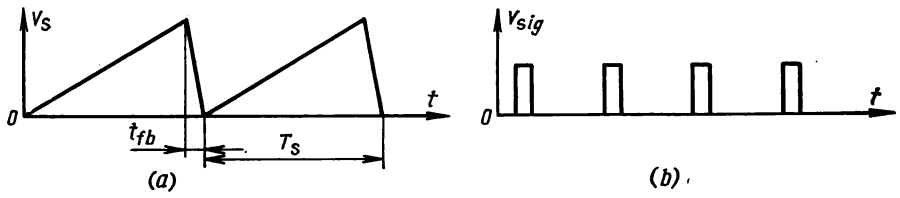


Fig. 7-19. Waveforms of voltages at the plates of an oscilloscopic CRT
(*a*) sweep voltage; (*b*) signal to be viewed

a horizontal line across the screen, moving at a uniform rate. The beam is said to *sweep* or *scan* the screen. At the end of a sweep, the sweep voltage falls to zero in time t_{fb}, and the beam is returned to the left-hand side before it starts out on its next sweep (this is known as *retrace*, the *flyback stroke*, or, simply, *flyback*). If we apply the signal representing a process to be viewed, for example, a pulse (Fig. 7-19*b*), to the vertical deflection plates, the beam will trace out the signal waveform because it is simultaneously driven by the sweep voltage horizontally across the screen and back, and is deflected vertically by variations in the applied voltage over the cycle. This pattern is a luminous time plot, or *oscillogram* of the applied signal. Generally, the rate of change of sawtooth voltage and, as consequence, the speed at which the beam sweeps across the screen can be varied at will. In this case, the sweep scale changes, too.

When CRT patterns are to be photographed, the brightness of the screen should be maximal, especially when the beam sweeps it at a high speed. Since it does not pay to raise the speed of electrons by increasing the voltage at the second anode (because this would bring down the tube sensitivity), use is made of postacceleration CRTs. They have additional intensifier electrodes made in the form of current-conducting rings with progressively rising potentials, put on the funnel-shaped portion of the tube (Fig. 7-20).

Sometimes, for example, when determining the phase or time shifts between, say, two signals, these signals must be made visible at the same time. For this purpose, *double-beam* or *dual-trace* oscilloscopic CRTs are employed. A double-beam tube has two electron guns arranged at such an angle that their axes intersect at the screen centre. With such an arrangement, one sweep trace may be superim-

posed on or set above the other by adjusting the direct voltage at one pair of the vertical deflection plates.

Radar CRTs. Somewhat different requirements are placed upon radar cathode-ray tubes. Here, the r.f. pulse reflected from an object (called the target) should be displayed on the screen so that the coordinates of the "pip" or "blip" on the screen will give the coordinates (range and bearing) of the target. The beam should be well focused to ensure the necessary accuracy in range and bearing. As a rule, use is made of magnetic cathode-ray tubes with long-persistence

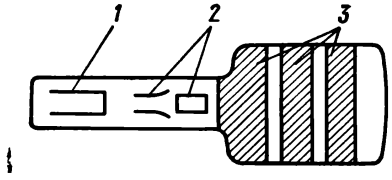


Fig. 7-20. Cathode-ray tube with post-acceleration of the beam
1—electron gun; *2*—deflection plates; *3*—helical intensifier anodes

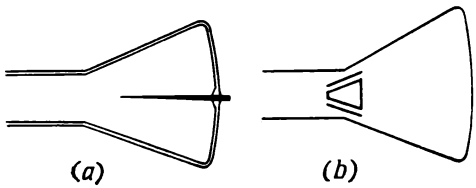


Fig. 7-21. Cathode-ray tubes with radial beam deflection
(*a*) with a rod electrode; (*b*) with cone electrodes

phosphors. Signals are displayed on the screen in polar coordinates. This is achieved by using a magnetic deflection structure consisting of a pair of coils. To the coils is applied a sawtooth current which sets up a linearly varying magnetic field deflecting the electron beam from the screen centre to the periphery along the radius. A special device makes the deflecting coils rotate around the tube neck, so that the beam traces out on the screen radial lines whose azimuth (direction) is changing slowly. The coils rotate in synchronism with a pencil-beam antenna which picks up echo signals from the object.

Normally, the tube is cut off ("blanked") by a negative modulating-grid voltage. An incoming echo signal develops a pulse which is applied to the modulating grid to render the tube conducting, and the screen displays a glowing spot. The distance between the light spot and the screen centre is proportional to the range of the target, and the angular position of the radius joining the screen centre and the "pip" corresponds to the bearing (or azimuth) of the target.

Radar sets also use CRTs with radial beam deflection. One of such tubes has a rod electrode sealed at the centre of the screen (Fig. 7-21*a*). The tube produces a circular sweep, for which purpose sine-wave voltages displaced in phase by 90° are applied to the two pairs of deflecting plates. The signal appearing at the central electrode deflects the electron beam so that it describes a circle around the screen periphery and deflects towards the centre. Radial beam deflection can also be produced by two cone electrodes (Fig. 7-21*b*).

In some cases, radar sets use *dark-trace* tubes. In contrast to conventional CRTs, dark-trace tubes have screens coated with a halide

of potassium. Crystals of potassium halide are normally transparent, but turn dark whenever fast electrons bombard them. As a result, the electron beam produces a dark magenta colour trace which is of long persistence and will remain on the screen for several days after the beam has been turned off. The phosphor can be discoloured by passing a current through the semitransparent film covering the phosphor.

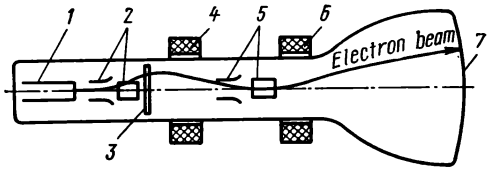


Fig. 7-22. Character-display tube (the Charactron)

1—electron gun; *2*—selection plates; *3*—matrix; *4*—convergence coil; *5*—reference plates; *6*—deflection yoke; *7*—screen

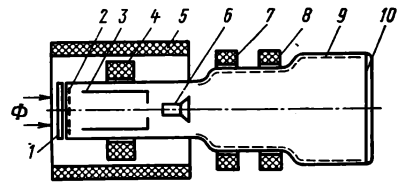


Fig. 7-23. Character-display tube (the Compositron)

1—matrix; *2*—photocathode; *3*—anode; *4*—selection coil; *5*—convergence coil; *6*—focusing electrode; *7*—magnetic lens; *8*—deflection yoke; *9*—conductive coating; *10*—screen

The pattern displayed by a dark-trace tube can be projected on a larger screen, using a strong light source and an optical projector.

For some time past, radar indicators have been employing *character-display* CRTs, that is, cathode-ray tubes displaying alphanumeric characters and other special symbols directly on their screens. The construction of one such tubes known as the Charactron (the trade name of General Dynamics/Electronics) is shown in Fig. 7-22. As is seen on leaving the electron gun, the focused electron beam 1 mm in diameter passes through the selection plates and founds itself at the matrix which is a plate with apertures made in the shape of characters or other symbols. The matrix has several hundred apertures about 0.5 mm in size each. Applying appropriate voltages to both pairs of selection plates, we may direct the electron beam to pass through the desired aperture in the matrix. On passing through the selected aperture, the beam assumes a cross section corresponding to the character in question. Further on, the beam is focused by the convergence coil, deflected by the reference plates towards the tube axis, and directed to an appropriate position on the screen by the magnetic field of the deflection yoke. The screen displays the symbol reproducing the shape of the matrix aperture through which the beam has passed. Because the matrix gives the beam an appropriate shape, this type of CRT is also known as a *shaped-beam* tube. The disadvantage of this tube is that the matrix cannot be replaced. This disadvantage is obviated in another type of character-display tubes, known as the *Compositron* (the trade name) where the matrix is put on the outside flat tube face opposite to the screen (Fig. 7-23). The inner surface is given a coat of a semitransparent photosensitive

material serving as a photocathode. When light strikes the photocathode through the matrix, the illuminated spots begin to emit electrons very intensively. As a result, the electron beam assumes the cross section of the illuminated spot on the photocathode. Here, the density of the beam corresponds to the shape of the matrix aperture. The magnetic selection coil deflects the electron beam in two directions at right angles to each other. In accordance with the direction and the value of deflection, only a part of the electron stream passes through the aperture disc in the anode. Thus, a narrow electron beam is formed past the anode, whose current density in the cross section corresponds to the shape of the matrix aperture in question. After that, the beam is focused by the convergence coil and directed to an appropriate position on the screen by the deflection yoke.

Television Picture Tubes (Kinescopes). These are cathode-ray tubes used in television receivers where electrical signals are translated into a visible picture on a luminescent screen. The requirements placed upon picture tubes depend on their purpose. The size of a picture and, as consequence, that of the screen should be convenient for observation. For man, most convenient angle of vision is about 15°. If the distance between an observer and a screen is a few metres, the optimum size of the picture ranges between 0.5 and 0.8 m. Present-day kinescopes having the screen with a diagonal dimension of 50 to 70 cm meet this requirement well.

The picture should be sufficiently bright and contrast. This can be achieved by using a rather powerful electron beam (up to several watts), phosphors with a high luminous efficiency (from 2 to 5 cd/W), and also an electron gun having a steep modulation characteristic.

Also important is picture definition, that is, image-detail visibility in a TV picture. Under the Soviet TV standard, each frame is formed by 625 scanning lines. With screens of about 0.5 m, this number of lines ensures an angular resolution* of about 1.5°. In the circumstances, the width of the luminous trace produced by the scanning beam (the spot diameter) should not exceed 0.3 to 0.5 mm. To ensure the necessary focusing and intensity modulation of the electron beam, present-day kinescopes use tetrode electron guns (Fig. 7-24*a*) with an electrostatic focusing system consisting of an immersion objective lens, an immersion projector lens, and a unipotential condenser lens.

For large beam deflection angles (90 to 110°), kinescopes employ magnetic deflection yokes. Large deflection angles make it possible to reduce the kinescope length, which is very important for a television receiver as a whole. A typical kinescope envelope is shown in Fig. 7-24*b*.

---

* Resolution is the ability of the picture-imaging medium to discern and reproduce two objects standing side by side on the televized scene.

Kinescope screens are coated with phosphors having white fluorescence and spectral characteristics as shown in Fig. 7-18*b*. As a rule, kinescopes use aluminized screens.

Colour kinescopes are of a special construction which is closely linked with the principles of colour television system in general; these picture tubes are outside the scope of the present book. Colour kinescopes, as well as various camera tubes, are discussed in a course on television.

Storage Tubes. These are tubes intended to write electric signals, store and read them out at a specified time.

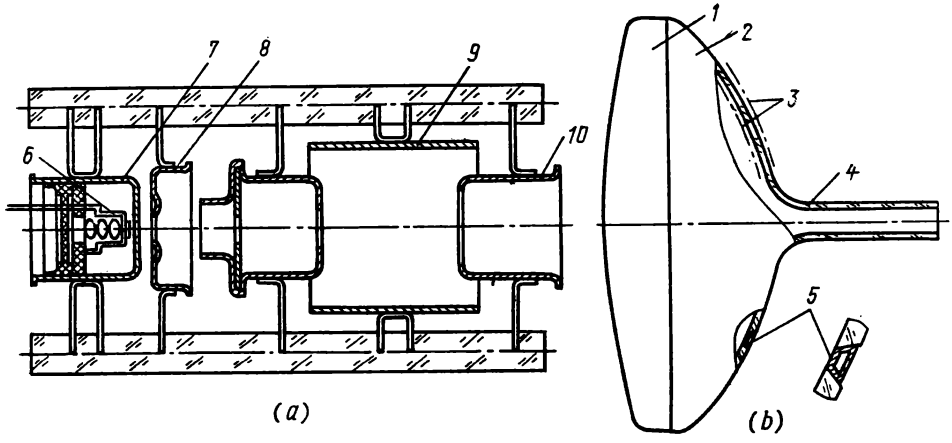


Fig. 7-24. Kinescope

(*a*) electron gun; (*b*) envelope; *1*—blank; *2*—funnel; *3*—conductive coating; *4*—neck; *5*—second anode lead; *6*—cathode; *7*—modulating grid; *8*—accelerating electrode; *9*—first anode; *10*—second anode

Storage tubes operate on the principle that the potential of a dielectric depends on the energy of electrons bombarding it. As has been shown in Sec. 7-5, the screen potential of a cathode-ray tube has two stable values, namely $V_s = 0$ at $V_{a2} \leqslant V'_{a2}$ and $V_s = V''_{a2}$ at $V_{a2} > V''_{a2}$ (Fig. 7-16). When the potential at the ultor (last anode) of the electron gun varies between V'_{a2} and V''_{a2}, the screen potential tends to V_{a2}. Thus, when the tube screen or a dielectric electrode, known in storage tubes as the *target*, is at one of its stable potentials, the target potential may be varied in synchronism with the modulating signal by modulating the velocity of beam electrons from V'_{a2} to V''_{a2}. If, at the same time, the beam is made to sweep the target, then the target potential will vary not only with time, but also with the coordinate x of the beam on the target. Consequently, as the beam sweeps the target, the potentials at the various spots on the target will vary in accordance with the shape of the modulating signal. As

a result, what is called a *charge image* or *pattern* will form on the target.

Methods and conditions of reproducing a charge image vary from tube to tube. We shall discuss only some of them.

One of the most commonly used storage tubes, the *barrier-grid-storage tube*, is shown in Fig. 7-25.

The barrier-grid tube has a conventional electron gun and deflecting plates (the electron beam may also be focused and deflected magnetically). The funnel-shaped part of the tube carries a target

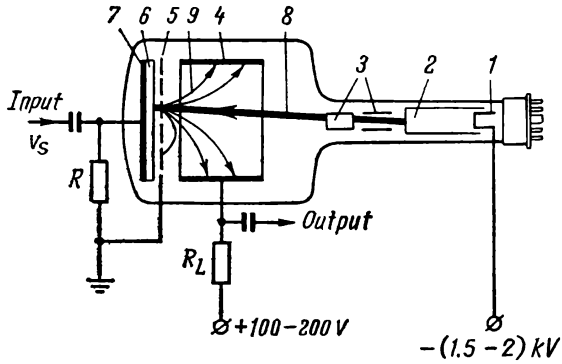


Fig. 7-25. Barrier-grid storage tube

1—cathode; *2*—electron gun; *3*—deflection plates; *4*—collector; *5*—barrier grid; *6*—target; *7*—signal plate; *8*—electron beam; *9*—secondary electrons

composed of a dielectric plate coated with a metal layer (the signal plate).

On the electron-gun side of the target, at a distance of 0.1 mm, is placed a very thin metal grid, known as the *barrier grid*. Next comes a circular electrode, called the *collector*, held at a very positive potential.

When no signal is applied, the beam traces out a horizontal line across the screen (beam sweep). The beam electrons strike the target and cause secondary emission. As this takes place, the target comes by an equilibrium potential close to the collector potential.

When a signal voltage is applied to the signal plate, the potential at the surface of the target which acts as a plate of a capacitor (made up of the signal plate, the dielectric and the target surface) begins to vary with time, following the signal voltage.

Let at time t_1 there be a positive signal voltage, v_{s1}. A rise in the signal plate potential brings about an increase in the voltage drop across resistor R, and the grid potential becomes negative with respect to the target. As a consequence, a smaller number of secondary electrons travel to the collector, the potential at the target element illuminated by the beam goes down, and the target element charges up to v_{s1}. When the beam leaves this target element, its potential continues to follow the input signal voltage but already at a new equilibrium level, v_{s1}.

In this way, the instantaneous signal voltage, v_{s1}, is written on a certain target element. Adjacent target elements will store other signal values written at other instants. Thus, the signal is analysed and stored on the target in the form of a charge pattern.

As the signal is written, it is simultaneously replicated across the resistor R_L in the collector circuit.

When the beam scans ("reads") the same line in the absence of a signal, the target elements discharge via the circuit made up of the

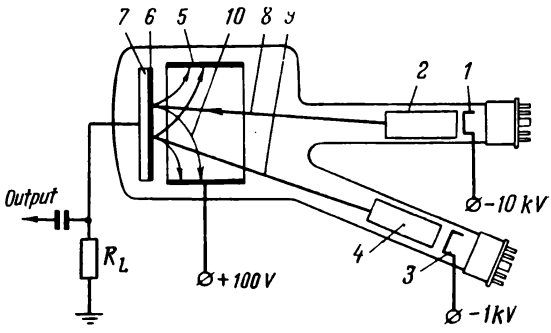


Fig. 7-26. Scan-conversion storage tube (the Graphechon)
1—cathode of the writing electron gun; *2*—writing e ectron gun; *3*—cathode of the reading electron gun; *4*—reading electron gun; *5*—collector; *6*—target; *7*—signal plate; *8*—writing beam; *9*—reading beam; *10*—secondary electrons

beam, ground and signal plate down to an equilibrium potential close to the collector voltage, and the collector current is modulated by variations in the charge image, that is, the signal. Thus, the signal is read out as the beam retraces the line. If reading is done with the signal turned off, the record (charge image) is erased.

If, while reading a previously stored signal, another signal different from the previous one is applied to the signal plate, the target elements will charge up to the voltage equal to the difference between the two potentials. Signal subtraction is said to take place, so such tubes are often called subtracting.

In addition to the above tube, use is also made of other storage tubes which sum (accumulate) signals, simultaneously write and display signals as waveforms, etc.

A special place among storage tubes is held by *scan-conversion* tubes, such as the Graphechon* (Fig. 7-26). Here, signals are written and read out by different electron beams each generated by a separate (writing or reading) electron gun. The reading gun makes an angle with the tube axis. The target is a very thin plate of magnesium fluoride, a dielectric in which a high volume conductivity can be induced by bombardment. On one side, the target is supported by a fine-mesh metal screen serving as a signal plate. An input signal is applied to the control electrode of the writing electron gun and intensity-modulates the electron beam. Because the writing gun operat-

* "Graphechon" is a proprietary name.— *Translator's note.*

es at a voltage of about 10 kV, its electrons travel to the target at a high velocity. As they penetrate into the dielectric, the electrons induce volume conductivity in the material. This bombardment-induced conductivity serves to discharge the target elements in proportion to the writing beam intensity, that is, in proportion to the signal voltage, thereby producing a charge image on the target surface.

In reading, the unmodulated beam scans the target and causes secondary emission from each target element in proportion to the stored potential. The secondary electrons are captured by the collector field whose voltage somewhat exceeds the equilibrium target potential. The collector current is thus modulated by the signal, and a voltage reproducing the signal is developed across resistor R_L.

Because the reading gun operates at a voltage of about 1 kV, its electrons have a considerably lower velocity. As a consequence, they do not penetrate into the target as deeply as the electrons of the writing beam. So, when the reading beam scans a line, the previously stored pattern is affected only to an insignificant depth. As a result, a signal written in a scan-conversion storage tube may be read out as many as several hundred times.

In some scan-conversion storage tubes, electron guns are arranged along the tube axis on either side of the target. The writing gun is placed on the signal plate side, because the signal plate is transparent to the writing beam electrons.

Storage tubes are used in radar sets to select targets against the background of clutter, and in moving target indicators. In the former case, a "stationary" signal is repeatedly written on the same target elements, so that the valid signal voltage is built up, while fortuitous interference cancels out. In the latter case, echoes from moving targets are built up, and those from stationary objects are cancelled out.

Storage tubes are also used as memory devices in computers. Data are coded as pulse sequences, written on the target, stored for a long time, and then read out when necessary to be used in computation.

Chapter Eight

GAS-FILLED TUBES

8-1. GENERAL

Definition. As their name implies, gas-filled tubes are electron devices which have some amount of gas or vapour introduced into them. Operation of these tubes is based on the electric discharge through the filling gas.

The envelopes of gaseous tubes (as they are also called) can be filled with inert gases (such as argon, neon or krypton), their mixtures, hydrogen or mercury vapour. The gas pressure in the envelope is low, being 10^{-1} to 10^{3} Pa. Gas-filled tubes are able to conduct much larger currents than vacuum tubes consuming the same power. This is because the former have a lower dynamic anode resistance and, as a consequence, a lower voltage drop between the anode and the cathode.

Gas-filled tubes are many and diverse. Some (such as *ignitrons*, *thyratrons* and *mercury rectifiers*) are used to rectify alternating current. Others are designed to stabilize direct voltages (for example, *stabilizer tubes*), operate as relays, switches (thyratrons and gas-filled spark-gap tubes), etc.

In the Soviet Union, the diagram symbols for gas-filled tubes are established by USSR State Standard GOST 7624-62.

8-2. ELECTRIC DISCHARGE IN GAS

In gas-filled tubes, electric charges are carried not only by electrons but also by positive ions. Free electrons and ions are produced through the ionization of gas atoms by high-velocity electrons, incident radiation, cosmic rays, etc. In addition, free electrons and ions may be given up by thermionic cathodes, photocathodes and also cold cathodes which emit electrons when struck by ions.

As an example, let us consider the physical processes taking place during an electric discharge in a gas-filled tube containing an anode and a cold (unheated) cathode (Fig. 8-1), supplied by a source, E, via a variable resistor, R.

Suppose that $R = \infty$ and the potential difference between the cathode and the anode is zero, $V_a = 0$. As already noted, the interelectrode space contains a small number of free electrons and ions produced through the ionization of the filling gas by incident light, cosmic rays and so on. Ionization is accompanied by the reverse process, known as recombination, which consists in the restoration of neutral atoms. These two processes are in dynamic equilibrium, and the number of charged particles remains constant. If the cathode is made of a material having a low work function, the same factors may force it to emit electrons as well. Some of the electrons which have escaped from the cathode fall back onto its surface; as a result, near the cathode, there is an electron cloud which is in dynamic equilibrium.

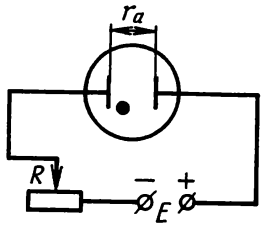


Fig. 8-1. Connection of a gas-filled tube in a circuit

Now let us assume that R is high but finite, and the voltage between the anode and the cathode is greater than zero. Under the action of V_a, a current due to the drift of electrons and ions flows in the interelectrode space. Since the mass of an ion is greater than that of an electron, ions travel in an electric field more slowly than electrons, and the ion current is lower than the electron current. The resultant current is the sum of the two currents because unlike charges move in opposite directions. Since more electrons than ions leave the interelectrode space per unit time, a positive space charge is developed in the tube.

As electrons travel to the anode and ions to the cathode, they may collide with gas molecules. The character of these collisions markedly depends on the velocity acquired by an electron along its free path, that is, on its energy at the instant of collision. If the electron velocity is low, collisions are elastic, and the energy of electrons remains the same. When electrons travel at a higher velocity, collisions may become nonelastic, the energy of at least one of the colliding particles may change, and the electrons of the atom may move to the excitation or ionization levels. The voltages at which electrons acquire the energy sufficient for these processes to be initiated are called the *excitation voltage*, V_e, and the *ionization voltage*, V_i, respectively. In the former case, the electron does not leave the atom but occupies a higher energy level, so that a lesser energy is required to separate the electron from the atom. As a rule, the excited atom is unstable, and, in a very short time, the electron returns to the original level giving up a quantum of radiant energy. The atom, however, may also be in a more stable excited state, known as the metastable state. The atom leaves this state only when it experiences a new collision; in this case, an excited electron may either escape from

the atom (ionization is said to take place) or regain its stable unexcited state. When ionization occurs, the electron leaves the atom, thereby forming a positive ion and a free electron. This is known as *volume ionization.*

Moving to the cathode and bombarding its surface, ions knock out free electrons, thus giving rise to *surface ionization.* In turn, the newly formed electrons bring about volume ionization, and so on.

The extent of volume and surface ionization is stated in terms of the factors α and γ defining the number of free electrons formed per electron or ion, respectively (with volume ionization, the number of free electrons formed per unit length of electron path).

Thus, when the potential difference between the tube electrodes is positive, the decrease in the number of charged particles constituting current through the tube is compensated for by volume and surface ionization.

When the wholly or partially ionized gas reaches a state in which the positive ions and negative electrons are roughly equal in number, we speak of the *plasma.*

Suppose that, under the action of some external factors, n_0 electrons left the cathode surface during a time t_1. In an electric field, the electrons moving toward the anode collide with gas molecules and ionize them. The newly formed free electrons join the primary n_0 electrons and also take part in volume ionization as they travel to the anode.

The ions formed in the course of volume ionization move toward the cathode and knock out new electrons which, in turn, participate in volume ionization. Thus, as they approach the anode, the number of electrons increases cumulatively. This phenomenon is called the *avalanche of charged particles.*

However, the current in the external circuit and, as a consequence, in the tube itself, is limited by the external resistance. At specified E and R, the tube reaches a steady state which the cumulative rise in the number of electrons is limited by the positive space charge formed near the cathode by ions.

In this steady state, the current density near the cathode is defined by an equation given in reference [1]

$$j_a = j_0 e^{\alpha r_a} / [1 - \gamma (e^{\alpha r_a} - 1)] \qquad (8\text{-}1)$$

where r_a is the distance between the cathode and the anode, and $j_0 = en_0$ is the density of the current caused by an external ionizing agency.

The type of gas discharge depends on the ratio of the number of electrons formed during the discharge to the number of primary electrons, defined as

$$M = \gamma (e^{\alpha r_a} - 1) \qquad (8\text{-}2)$$

Electric Discharges in Gases. If M is below unity, the anode current density, in accordance with Eq. (8-1), depends on j_0. When the external ionizing agency ceases to affect the tube, the discharge ceases, too. Here, we have a *nonself-maintained discharge*. With this form of discharge, as E and, as a consequence, the electric field in the tube increase, α and γ change so that the current density j_0 rises.

If $M \geqslant 1$, that is, each electron leaving the cathode due to all types of ionization gives rise to at least one new electron, the discharge is maintained even after the external ionizing agency ceases to affect the tube. This is a *self-maintained discharge*.

The intensity of ionization depends on the kind of gas, its pressure and temperature inside the envelope, and also the potential difference between the electrodes. If all factors, except the potential difference, remain unchanged, the discharge processes in the interelectrode space proceed according to the discharge volt-ampere characteristic shown in Fig. 8-2.

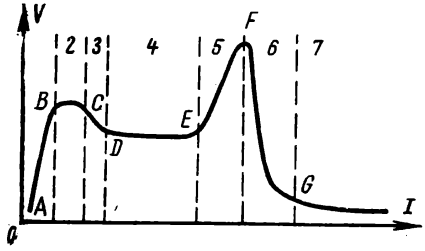


Fig. 8-2. Volt-ampere characteristic of an electric discharge in gas

Here, the independent variable (laid off as abscissa) is the current through the tube, because the potential difference between the electrodes depends on the state of the ionized gas, that is, on the motion of charged particles and the distribution of space charges between the anode and the cathode.

Region *1* on the volt-ampere characteristic represents a nonself-maintained discharge. The transition from point A to point B illustrates the processes taking place when R is reduced (Fig. 8-1); for all points of this portion, the number of electrons leaving the cathode is small, gas ionization is insignificant (α and γ are small) and M is below unity. The resistance is very high, and a negligible current flows through the tube.

An increase in current brings about a self-maintained discharge ($M = 1$). Region *2* represents the initial stage of the self-maintained discharge, known as the *Townsend* or *dark discharge*. This also occurs at a low current limited by the high external resistance.

As the current rises further, the Townsend discharge transforms to the *glow discharge* (region *4*), that is the gas begins to glow because excited electrons returning to stationary states radiate energy as photons.

In region *3*, the voltage across the tube falls. Point C in this region reflects the more intensive ionization at a cathode area where emission is maximal. Since ions move more slowly than electrons, a large number of ions neutralizing the negative space charge due to elect-

rons and facilitating their travel toward the anode accumulate at this spot, thereby forming a column of the more ionized gas. The column increases in cross section with increasing current, and, as it reaches a certain size, the directed flow of electrons toward the anode and volume ionization begin to prevail over the diffusion of charged particles into the surrounding gas. Now, the discharge can be maintained without so high a potential difference between the cathode and the anode, and the voltage across the tube decreases (portion *CD*).

Under conditions of the glow discharge (region *4*), the above processes become more stable. Near the cathode appears a positive space charge changing the potential distribution. This charge occupies what is known as the cathode-fall space. At first (near point *D* in region *4*), the cathode-fall space covers a small cathode area. Then, the cathode area subjected to the cathode fall grows and the current through the tube increases.

As this takes place, the ionized gas column expands in proportion and, as a consequence, the current density and the potential difference between the electrodes remain the same as before. This constancy of voltage with changing current through the tube, distinctive of the glow discharge, forms the basis for stabilizer tube operation. When the cathode fall extends to cover the entire cathode surface, new charged particles are required to step up the current through the tube. This may be done only by raising the potential difference between the electrodes, which causes a rise in volume and surface ionization. This discharge (region *5*) is called the *abnormal glow*.

The voltage at point *F* is so high that the velocity of ions greatly increases. The cathode bombarded by ions grows hot and begins to emit electrons. In mercury-pool tubes, a rise in cathode temperature results in a more intensive mercury evaporation. The density of mercury vapour builds up, and electrons begin to collide with mercury molecules more frequently thereby forming an increasing number of ions. Ions being very close to the cathode surface make up a very strong electric field (about 10^6 to 10^8 V/cm) which brings about field emission.

Thermionic and field emissions sharply raise the number of electrons in the tube; their space charge neutralizes the positive space charge due to ions and the voltage goes down (region *6*). An *arc discharge* is said to occur (region *7*). A distinction of the arc discharge is a small voltage drop between the cathode and the anode (mainly due to an insignificant cathode fall) at a heavy current. The arc discharge is utilized in high-power gas-filled tubes, mercury-vapour rectifier diodes and ignitrons.

In hot-cathode tubes, the arc discharge occurs at low electrode voltages because the cathode is heated by an external voltage source rather than by ions. Such discharge known as the *low-voltage arc*

is used in gas-filled thermionic diodes and thyratrons. The low-voltage arc may occur at anode potential smaller than the ionization potential or even the excitation potential. In the former case, the low-voltage arc is called normal, in the latter, abnormal.

Gas-filled tubes also use the *high-frequency discharge* which takes place due to the action of a high-frequency electromagnetic field. If we disconnect the electrodes of the tube of Fig. 8-1 from the power source and use them as the plates of a capacitor in a high-frequency resonant circuit, a high-frequency field will be set up in the interelectrode space and a high-frequency discharge will occur. A distinction of this discharge is that heavy ions do not have time to reverse their motion as the potentials at the electrodes change polarity, and form a positive space charge. The high-frequency discharge may occur at low potential differences between the electrodes.

This discharge is used in radar resonant spark-gap tubes.

8-3. ARC-DISCHARGE TUBES

Arc-discharge tubes include hot-cathode tubes with a nonself-maintained discharge (thyratrons) and also tubes with a self-maintained discharge (such as ignitrons and mercury-vapour rectifier diodes).

The Thyratron. This is a gas-filled tube containing an anode, a cathode and one or more grids. Most frequently use is made of hot-cathode thyratrons utilizing the nonself-maintained arc-discharge. The construction of a thyratron containing one grid is shown in Fig. 8-3. The envelopes of low-voltage thyratrons are filled with inert gases, and those of high-voltage and pulse thyratrons are filled with mercury vapour and hydrogen, respectively. Thyratron grids are usually nickel or molybdenum discs with apertures of various shapes and sizes.

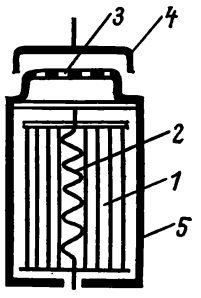


Fig. 8-3. Thyratron
1—cathode; *2*—heater; *3*—grid; *4*—anode; *5*—screen

By varying the grid voltage, we can change the electric field strength and control the instant at which the tube can fire. The grid voltage at which the discharge occurs also depends on the anode voltage.

Figure 8-4*a* shows the relation $I_a = \psi(V_g)$ plotted for different values of V_a. To measure this characteristic, the anode voltage, V_a, is first set to zero, then a high negative voltage, V_g, is applied to the grid, and the anode voltage is raised to a specified value. So long as the grid is considerably negative, the current through the tube is equal to zero. If the grid is made less negative, the fastest electrons will overcome the retarding field, pass through the grid and move

toward the anode. At a certain grid voltage, $-V_g = -V_{g,f}$, the filling gas is ionized and the current suddenly rises. The thyratron is said to fire. Further variations in the grid potential in either direction from $-V_{g,f}$ do not affect I_a because the grid loses control. This happens because the field developed by the positive ions formed during the discharge and attracted to the grid neutralizes the field of the negative grid. The discharge may be stopped only by considerably reducing or turning off the anode voltage.

If we raise the anode potential ($V'_a > V_a$), the discharge will occur at a more negative grid voltage, $-V'_{g,f}$; the current will be

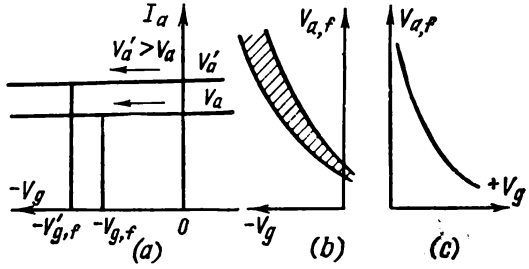


Fig. 8-4. Thyratron characteristics
(a $I_a = \psi(V_g)$; (b) "negative" control characteristic; (c) "positive" control characteristic

higher because for faster electrons the gas has a higher ionization factor.

The curve relating the firing potential of a thyratron to its anode voltage is called the *control characteristic of the thyratron* (Fig. 8-4*b* and *c*). The control characteristic of one and the same thyratron operating on an alternating voltage is anything but stable. If plotted several times, the curves will lie within a certain firing or conducting region (the shaded area in the figure). In this they differ from the static control characteristics plotted for d.c. operation. The existence of a firing region can be accounted for by a prefiring grid current which is maintained by delay in the deionization of the thyratron when it is cut off and, in addition, by emission from the grid. Within the thyratron the prefiring current flows from the cathode to the grid. The grid circuit resistance is also responsible for the shift in the control characteristic. The ions surrounding the grid induce a current which is directed away from the grid in the external circuit. When the grid draws current, the voltage drop across the grid circuit resistance makes the grid less negative and shifts the control characteristic.

When the anode is at a sufficiently high potential, a self-maintained glow discharge may occur between the grid and the anode of the thyratron. As a precaution, data sheets specify maximum safe anode voltages for all types of thyratrons. This undesirable pheno-

menon may be avoided by using another grid, known as the screen grid, which also helps to increase the maximum safe anode voltage. The screen grid also makes it possible to expand the range over which the thyratron current may be controlled. The construction of a screen-grid thyratron is shown in Fig. 8-5, and its control characteristics plotted for various V_{g2} appear in Fig. 8-6.

Sometimes, thyratrons use grids made of closely spaced fine molybdenum or nickel wire (such as the Tacytron). The grid placed

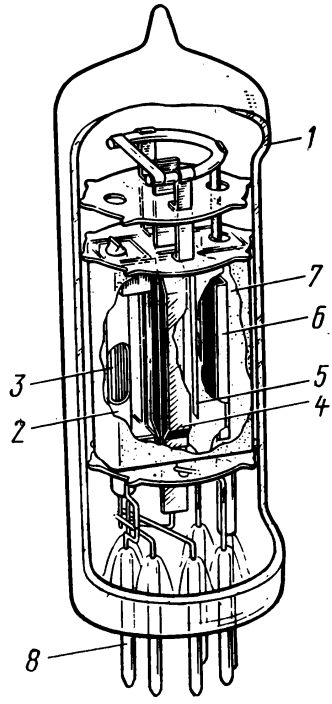


Fig. 8-5. Screen-grid thyratron
1—envelope; *2*—cathode; *3*—heater; *4*—control grid; *5*—screen grid; *6*—anode; *7*—screen; *8*—pins

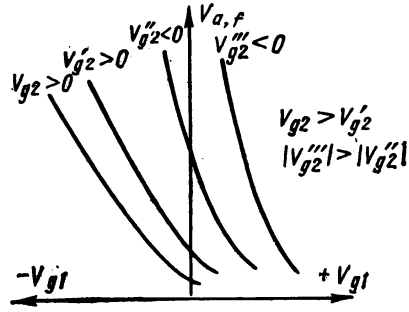


Fig. 8-6. Control characteristics of the screen-grid thyratron

inside the screen enclosing the anode and the cathode divides the interelectrode space into the anode and cathode sections. The lines of force from the anode do not practically penetrate through the closely spaced grid wires into the cathode section, so a dense negative space charge is formed near the cathode. Electrons diffusing between the grid wires travel to the anode at a low velocity, so no ionization takes place in this area. The discharge occurs only in the anode section where electron velocities are sufficiently high (at small distances from the grid, the potential is practically equal to the anode potential). In these tubes, grid voltage can be used to control the instants at which the tube both fires and is extinguished.

Thyratrons have found application in electronic relays whose control circuits require low currents to turn on and off circuits handling heavy currents, and also in rectifier circuits. In the latter,

the value of rectified current may easily be controlled by varying the grid voltage. The thyratron fires during the positive half-cycles of anode voltage at the instant depending on the specified grid voltage. After the tube has fired, anode voltage decreases to the value of the voltage drop existing in an arc discharge (15 to 20 V). The thyratron conducts current starting from the instant it fires till the end of the positive half-cycle. The duration of the anode current pulse and, as a consequence, its average value may be changed by varying the grid voltage and controlling the instant at which the tube fires.

The self-maintained arc discharge is utilized by high-power mercury-vapour rectifier diodes and ignitrons which are employed in high-power rectifiers practically not used in electronic circuits.

8-4. GLOW- AND HIGH-FREQUENCY-DISCHARGE TUBES

The glow discharge is utilized in gas-filled stabilizer tubes and also in low-power pulse switches called glow thyratrons.

The Stabilizer Tube. The construction of a stabilizer tube is shown in Fig. 8-7. Its cathode is a cylinder enclosing the rod anode. The inner surface of the cylinder cathode is given a coat of an activator.

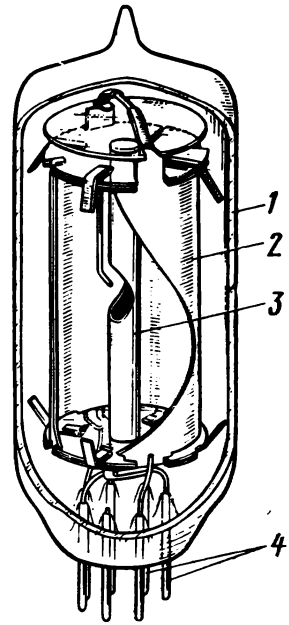


Fig. 8-7. Stabilizer tube
1—envelope; *2*—cathode; *3*—anode; *4*—pins

The stabilizer tube is connected in a circuit as shown in Fig. 8-8. The voltage which the stabilizer tube maintains constant is called the *maintaining* (or *stabilization*) *voltage*, V_m (Fig. 8-9), and the currents representing the limits of the normal glow region are termed the maximum, I_{max}, and minimum, I_{min}, currents.

The series section of the circuit contains a ballast or current-limiting resistor R_s. Its value is chosen so that at nominal V_0 and R_L the current through the tube will be I_{av}. When the supply voltage V_0 or the value of the load resistor are changed, the current through the tube also changes, but the voltage across it remains the same, and, as a consequence, so does the voltage across the load resistor connected in parallel with the tube. The maintaining voltage mainly depends on the work function of the cathode material and the ionization potential of the filling gas.

The Glow Thyratron. This finds use in counting and pulse circuits. The construction of two Soviet-made glow thyratrons is shown

in Fig. 8-10. The miniature envelope is filled with neon at a pressure of 10^3 Pa. In the MTX-90 triode thyratron (Fig. 8-10*a*), the cathode

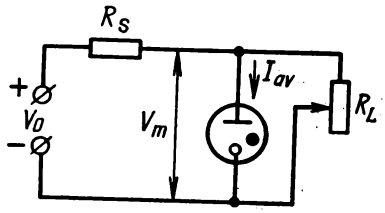


Fig. 8-8. Connection of the stabilizer tube in a circuit

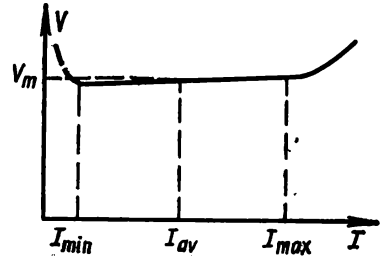


Fig. 8-9. Volt-ampere characteristic of the stabilizer tube

is a nickel cylinder coated with cesium on the inside. The anode is the end face of a piece of molybdenum wire; the wire itself is

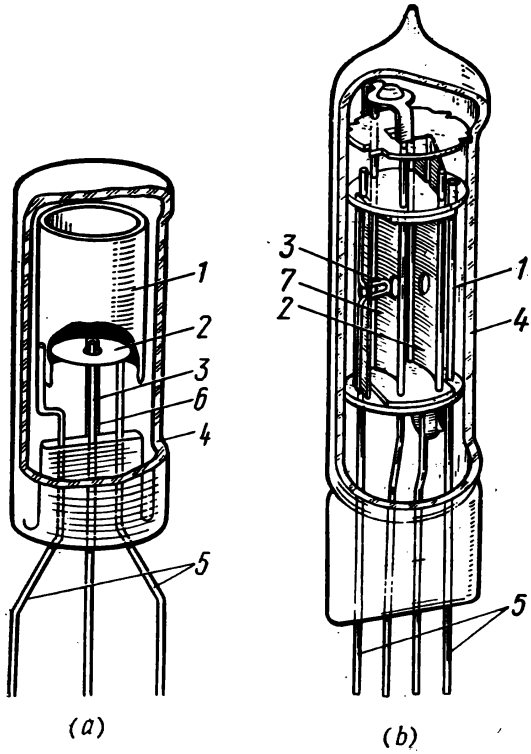


Fig. 8-10. Glow thyratrons
(*a*) Soviet-made type MTX-90 triode thyratron; (*b*) Soviet-made type TX-3Б screen-grid thyratron; *1*—cathode; *2*—grid; *3*—anode; *4*—envelope; *5*—pins; *6*—anode glass jacket; *7*—screen grid

enclosed in a glass jacket near the end. The grid protects the anode against direct exposure to charged particles moving from the cathode. The type TX-3Б double-grid (or tetrode) thyratron has a rod cathode, and its grids are plates with round apertures.

Figure 8-11 shows the main characteristics of glow thyratrons. In the triode thyratron, the firing voltage depends on grid current. That is why they are often called *current controlled thyratrons*. Here, the grid is held at a low potential which causes an auxiliary trigger discharge between the grid and the cathode which helps to reach

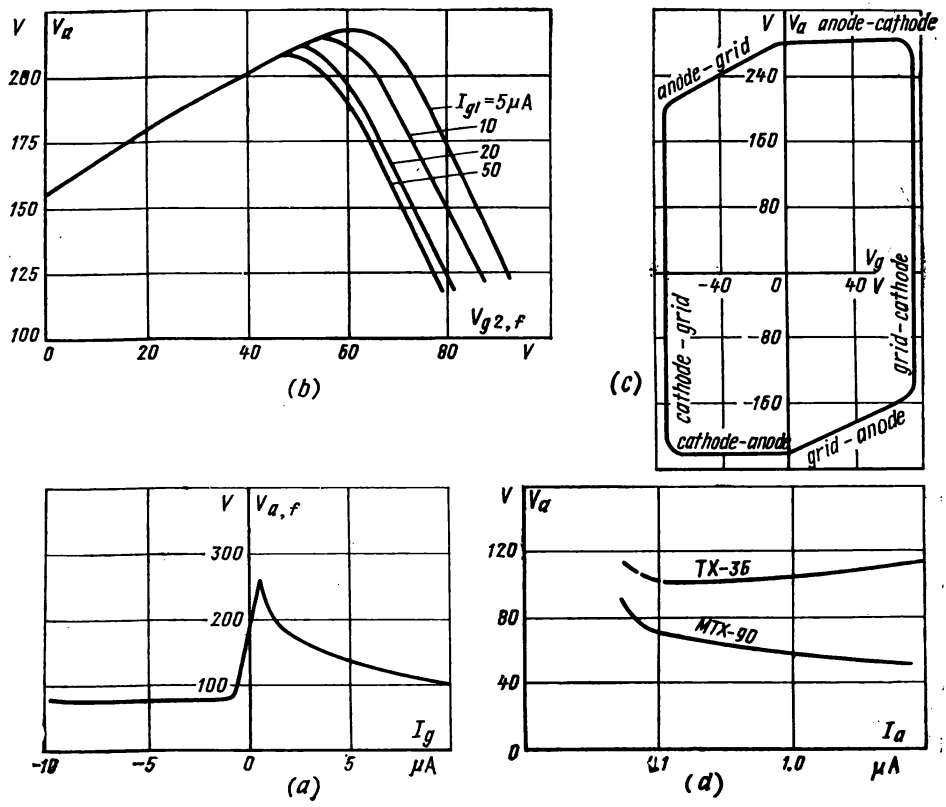


Fig. 8-11. Characteristics of glow thyratrons

(*a*) firing characteristic of the Soviet-made type MTX-90 thyratron; (*b*) firing characteristic of the Soviet-made type TX-3Б thyratron; (*c*) breakdown characteristic; (*d*) volt-ampere characteristics

a glow discharge between the grid and the anode. The best results are obtained with low grid currents as can be seen from the drooping section of the glow thyratron firing characteristic (Fig. 8-11*a*).

In the tetrode thyratron, the instant at which the tube fires is controlled by varying the potential between the first and the second grids (this is known as the *potential control*). The first grid is held at a positive potential of 50 to 100 V which ensures a preliminary trigger discharge between the cathode and the first grid. The second grid is less positive than the first. As V_{g2} rises, the retarding field between the grids becomes weaker, electrons move toward the anode and ionize the gas between the second grid and the anode thereby

causing a glow discharge. Referring to the firing characteristics of this thyratron (Fig. 8-11*b*), the instant at which the tube fires can be controlled not only by varying voltage but also the first-grid current, I_{g1}.

Sometimes, thyratrons may operate with their grids and anodes held at potentials of opposite polarities. In this case, it is convenient to use what is known as the *breakdown characteristic* (Fig. 8-11*c*). Referring to the figure, the glow thyratron fires at any voltage within the area bounded by the characteristic. The inscriptions in each quadrant indicate the space in the tube where the discharge occurs.

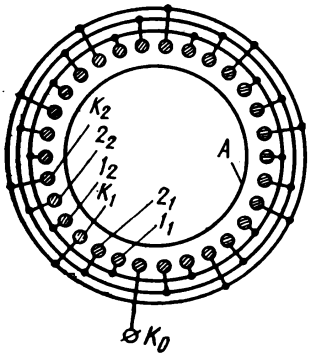


Fig. 8-12. Decade-counting tube

K_0—zero cathode *1*—first transfer or guide electrodes; *2*—second transfer or guide electrodes; K_1, K_2—index cathodes; *A*—anode

The volt-ampere characteristics of glow thyratrons (Fig. 8-11*d*) demonstrate the stabilizing ability of these tubes, distinctive of the glow discharge in general.

Glow thyratrons are intended for operation in pulse circuits. So, the time of de-ionization after anode voltage has been turned off should be short.

The Decade Counting Tube. This is a multi-electrode glow-discharge switch.

Figure 8-12 shows the construction of a double-pulse decade counter. It consists of a central anode surrounded by three groups of cathodes, namely main or index cathodes K, first transfer or guide electrodes marked '1' and second transfer or guide electrodes labelled '2', connected by cathode rings. The main index cathode called the zero cathode K_0 has a separate terminal.

The anode is maintained positive with respect to the index cathodes; a positive voltage of about 50 V is also applied to the first transfer electrodes. In the quiescent state, the glow discharge is maintained only between the anode and the zero cathode. The discharge is transferred to other electrodes when two negative series guide pulses with an amplitude of about 100 V and a duration of 20 to 30μs are applied to the tube.

The first guide pulse arrives in the circuit of the first transfer electrodes and transfers the glow discharge to the guide electrode labelled 1_1. The second guide pulse arriving in the circuit of the second transfer electrodes transfers the discharge to the nearest guide electrode 2_1. When the second guide pulse ceases, the discharge transfers onto the first main or index cathode K_1, and so on. The pulses are counted by glows at the index cathodes, visible through the tube glass. The discharge is transferred to the zero cathode by breaking the index-cathode circuit.

A distinction of double-pulse decade-counter tubes is that the direction of discharge transfer can be reversed if the guide pulses are used the other way round. Then, any number of pulses can be subtracted from those already counted.

There are also decade-counter tubes operating with a single control pulse.

Resonant Spark-Gap Tubes. Resonant spark-gap tubes (Fig. 8-13) utilize the high-frequency discharge.

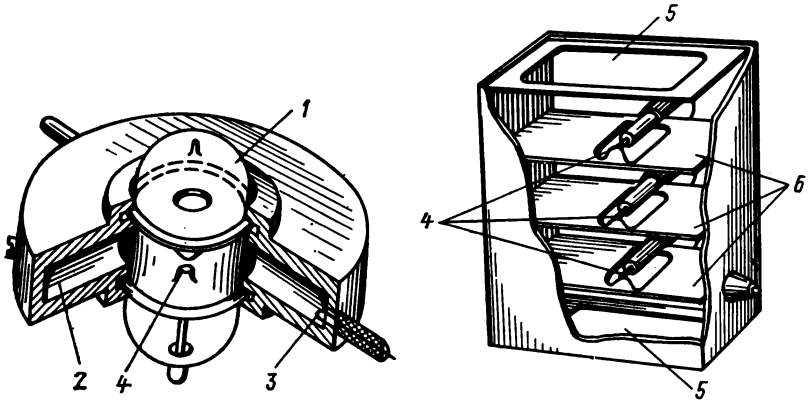


Fig. 8-13. (*a*) Narrow-band coaxial spark-gap tube and (*b*) wide-band waveguide spark-gap tube

1—glass envelope; *2*—cavity resonator; *3*—coupling loop; *4*—spark gap; *5*—window coupling the spark gap to the waveguide; *6*—aperture diaphragms

A coaxial spark-gap tube (Fig. 8-13*a*) contains two cone-shaped electrodes. The space between their apexes serves as a spark gap. The disc leads of the electrodes are connected to a cavity resonator and make up its capacitance.

A waveguide spark-gap tube (Fig. 8-13*b*) is a parallelepiped whose dimensions correspond to those of a waveguide. The glass windows sealed into the spark-gap tube walls couple the spark-gap with the waveguide. Inside, the parallelepiped is divided by diaphragms with resonant apertures. Thus, the tube is a system of coupled resonant circuits, so its bandwidth is sufficiently large.

To facilitate the discharge, use is sometimes made of additional trigger electrodes which cause initial gas ionization.

Most radar sets use one and the same antenna to transmit and receive signals. Signals reflected from a target are received during the intervals between the transmitted pulses. The power of transmitted pulses reaches several megawatts. Obviously, if such a signal arrived at the receiver mixer, the latter would be damaged. So, to avoid this, the receiver input contains a gas-filled spark-gap tube (Fig. 8-14). During the intervals between the strong transmitted

pulses, the spark gap serves as a part of a cavity resonator tuned to the frequency of the incoming signal. During the transmitted pulse, the resonator produces high-power oscillations. The cavity formed by the spark gap electrodes generates a high voltage causing a high-frequency discharge. Since the resistance of the tube during the discharge is very low, it practically shorts out the resonant circuit. The length of the high-frequency line connecting the spark

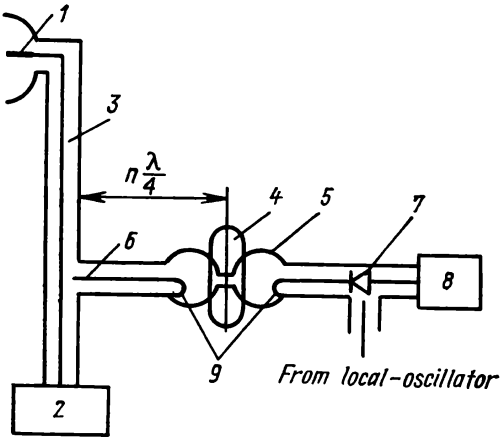


Fig. 8-14. Connection of a resonant spark-gap tube in a circuit

1—antenna; *2*—transmitter; *3*—antenna feeder line; *4*—spark gap; *5*—cavity resonator; *6*—coupling rod; *7*—crystal diode; *8*—amplifier; *9*—coupling loop; $n = 1, 3, 5 \ldots$

gap to the antenna is equal to a quarter of the wavelength (or to an odd number of quarters), so its input impedance at the point where it is connected to the antenna line is very high and the transmitted signal breaking through into the receiver mixer is reduced to a fraction of its original value.

Part Two

SEMICONDUCTOR DEVICES

Chapter Nine

A GENERAL OUTLINE AND PHYSICAL PRINCIPLES OF SEMICONDUCTOR DEVICES

9-1. CLASSIFICATION AND DESIGN OF SEMICONDUCTOR DEVICES

Definition. As its name implies, a semiconductor device is an electron device in which the characteristic distinguishing electronic conduction takes place within a semiconductor.

It is customary to assume that semiconductors are solids or liquids whose resistivity at room temperature ranges from 10^{-4} to 10^{10} ohm cm. Materials with $\rho \leqslant 10^{-4}$ ohm cm are called conductors (metals) and those with $\rho \geqslant 10^{10}$ ohm cm are termed insulators (dielectrics). The division of materials into these three groups is rather relative. It is especially difficult to draw a distinction between semiconductors and dielectrics because their energy structures have much in common, in particular, both have energy gaps separating valence from conduction bands. The difference in resistivity depends only on the width of the energy gap. As a rule, semiconductors include materials with a rather narrow energy gap (not more than 3 eV).

The distinction between conductors (metals) and semiconductors is more pronounced. The resistivity of metals increases approximately linearly with temperature

$$\rho = (\text{approx.})\ \rho_0 T/T_0 \tag{9-1}$$

where ρ_0 is the resistivity at $T_0 = 300$ K.

In contrast, the resistivity of semiconductors decreases with increasing temperature

$$\rho = \rho_0 \exp(\beta/T) \tag{9-2}$$

where β is a positive quantity.

For most metals, variations in ρ are about 0.5% per degree centigrade, and for semiconductors, 5% or more per degree centigrade.

Semiconductors include a large number of elements mainly from Groups IV through VI of Mendeleev's periodical system (such as C, Si, Ge, As, P, Se and Te) and also a considerable number of binary and triple compounds of elements from Groups I through VIII, such

as $A^{I}B^{VII}$, $A^{I}B^{VI}$, $A^{I}B^{V}$, $A^{II}B^{VI}$, $A^{III}B^{V}$, $A^{IV}B^{VI}$, $A^{I}B^{III}B_2^{VI}$ and $A^{I}B^{V}B_2^{VI}$.

Not all of these materials, however, are used in the manufacture of semiconductor devices. The basic materials are germanium (Ge) and silicon (Si). For some time past, gallium arsenide (GaAs) which is a type $A^{III}B^{V}$ compound has also been finding application.

The main electric and physical characteristics of these materials are given in Table 9-1.

Table 9-1

Characteristic	Germanium	Silicon	Gallium Arsenide
Nucleus charge	32	14	—
Valence S	4	4	—
Permittivity ε, relative units	16	12	11
Melting point T_m, °C	940	1420	1280
Effective mass of electrons m_n^*	0.22	0.33	0.07
Effective mass of holes m_p^*	0.39	0.55	0.5
Width of energy gap ** ΔE_g, eV	0.67	1.11	1.4
Electron mobility μ_n, $cm^2/(V\ s)$	3800	1400	8500
Hole mobility μ_p, $cm^2/(V\ s)$	1800	500	450
Intrinsic resistivity ρ_i, ohms cm	50	2×10^5	4×10^8
Intrinsic concentration n_i, cm^{-3}	2.5×10^{13}	2×10^{10}	1.5×10^6
Diffusion constant of electrons D_n, cm^2/s	100	36	290
Diffusion constant of holes D_p, cm^2/s	45	13	12

* The effective mass is given in terms of the rest mass of an electron at rest m = (approx.) 9.1×10^{-31} kg.

** All subsequent parameters are given at room temperature T = (approx.) 300 K

Design of Semiconductor Devices. The operating volume of a semiconductor device is limited by the physical boundaries of a single crystal of a semiconductor material. As is known from a course on physics, in contrast to a polycrystal, a single crystal has a unitary crystal lattice.

Most semiconductor devices use single crystals nonuniform in their electric and physical properties. In such crystals, we may distinguish regions where the population of electrons prevails over that of holes or vice versa, or regions with different resistivity.

That part of a crystal where such regions border upon each other is called the *P-N transition region.* The physical processes taking

place here underlie the operation of most semiconductor devices. *P-N* transition regions in semiconductors devices are made by fusing together two single crystals of the same material but containing different impurities, by diffusing impurities into different regions of

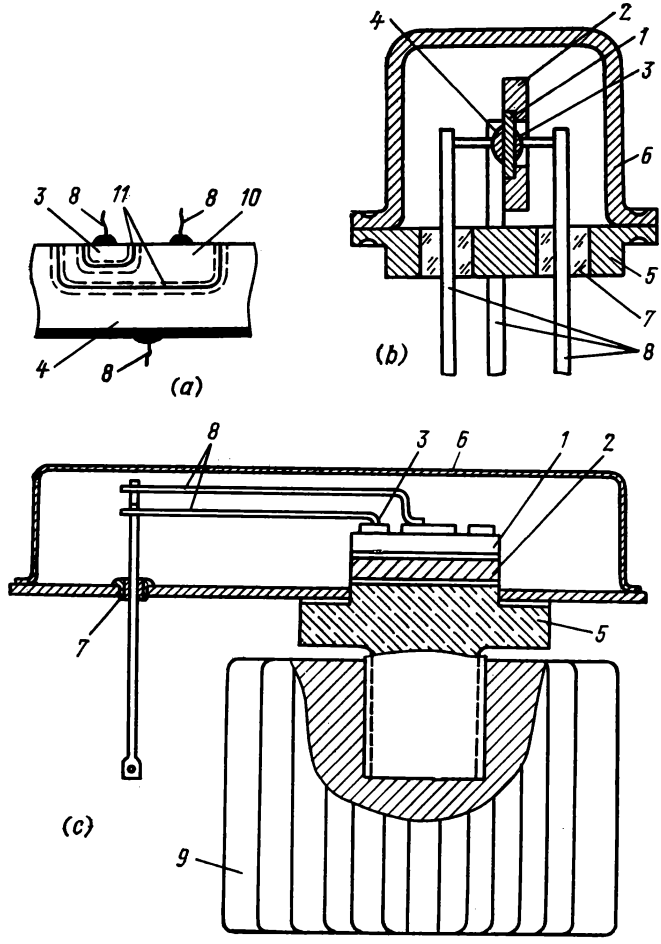


Fig. 9-1. Construction of transistors

(*a*) a single crystal of silicon with the electrodes of a planar transistor; (*b*) an alloy transistor in a metal case; (*c*) a high-power transistor with a collector heat sink; *1*—semiconductor wafer; *2*—crystal support; *3*—emitter; *4*—collector; *5*—transistor stem; *6*—case; *7*—insulator; *8*—leads; *9*—heat sink; *10*—base region; *11*—*P-N* junction

a single crystal, by fusing together a single crystal of a semiconductor and a metal, or by fusing together single crystals of two different semiconductors. *P-N* transition regions and physical processes in them are discussed in more detail in Chapter 10.

The surface of a single crystal near *P-N* transition regions is coated with a layer of a metal serving as an electrode to which thin-wire leads are then welded.

Figure 9-1*a* shows a section of a silicon crystal used in a *bipolar transistor* which is one of the most common semiconductor devices. The substrate of the transistor is a silicon crystal having an increased concentration of holes (*p*-Si). This part of the crystal is the *collector region*. Another part of the crystal has an increased population of electrons (*n*-Si); this is the *base region*. Inside the base, there is the *emitter region* with an increased concentration of holes (*p*-Si). The boundaries between each pair of regions make up *P-N* transition regions (bounded by the dashed lines in the figure).

In most cases, semiconductor devices are enclosed in airtight metal, glass or plastic envelopes or cases isolating their crystals from the atmosphere (Fig. 9-1*b*). The electrodes of a semiconductor device are welded to fine-wire leads which are connected to power sources and other electronic circuit components.

In low-power devices, crystals are tenths or hundredths of a cubic millimetre in size; in high-power devices, they may be from unity to several tens of cubic millimetres large. The size of a case depends on the purpose of the crystal device and its maximum power dissipation. Low-power devices are miniature, they do not exceed a few millimetres in size; high-power devices may be as large as several centimetres; sometimes, they are equipped with heat sinks (Fig. 9-1*c*).

The last few years have seen wide application of unpackaged devices (without cases), in which crystals are covered by protective films. Such devices used in hybrid thin-film integrated circuits are even smaller than packaged crystal devices (enclosed in cases).

Basic Classification. Semiconductor devices are many and diverse and may be classed in several ways.

Most commonly, semiconductor devices are classified according to the manner in which they convert energy. In this classification, they may be divided into *electric*, *photoelectric*, *radiation*, *thermoelectric*, *tensoelectric* and some other types. The most numerous are electric conversion semiconductor devices. These include a majority of *crystal diodes*, all types of *transistors*, and *thyristors*.

The progress made in the field of infrared equipment, lasers and optoelectronics has necessitated expansion in photoelectric and radiation devices (such as *photoresistors*, *photodiodes*, *phototriodes*, *crystal photovoltaic cells*, *solar batteries* and *light-emitting diodes*).

The other types of semiconductor devices are not so numerous, although such physical phenomena as the *piezoelectric effect*, the *Hall effect* and the *Seebeck effect* are finding an increasing application in semiconductor transducers, coolers, etc.

In addition, semiconductor devices may be classed by the semiconductor materials they use (for example, *silicon*, *germanium* or

gallium arsenide devices), operating frequency (audio-frequency, radio-frequency and microwave devices), their maximum power dissipation, design (*alloy-junction*, *diffusion*, *mesa transistors*, *P-N-P* and *N-P-N* devices), and so on.

In the Soviet Union, the nomenclature and definitions for semiconductor devices are set up by USSR State Standard GOST 15133-69. The diagram symbols for them are established by GOST 2.730-73 and GOST 10862-72.

9-2. ENERGY DIAGRAMS OF SEMICONDUCTORS

As is known from a course on physics, the energy spectrum of solids consists of several bands. The energy levels of atoms forming a crystal structure are split into several bands separated from one another by forbidden bands or energy gaps which cannot be occupied

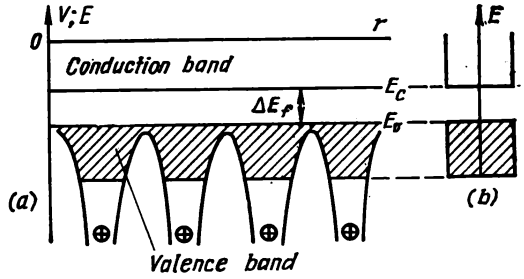


Fig. 9-2. (*a*) Potential and (*b*) energy diagrams of an intrinsic semiconductor

by particles. The electric and physical properties of a solid depend on the location of two top energy bands formed due to the splitting of the energy levels of valence electrons (the valence band) and excitation and ionization levels of electrons in atoms (the conduction band). In the energy diagrams of semiconducting materials, these two bands are separated by a relatively narrow (not more than 3 eV) energy gap. In Fig. 9-2*a*, these bands are shown on the V (r) plot of a crystal of an intrinsic semiconductor, and in Fig. 9-2*b*, as a one-dimensional energy-level diagram.

Semiconductor devices seldom use intrinsic semiconductors. As a rule, they employ semiconductor materials with predominating electron (or negative, N-type) conduction or hole (positive or P-type) conduction. For this purpose, certain impurities are added to a purified semiconductor (the material is said to be alloyed).

Local Impurity Levels. For the tetravalent germanium and silicon, the donor impurities, that is, the impurities which establish electron conduction, are elements of the fourth group of Mendeleev's periodic table (such as P, As and Sb), and the acceptor impurities establishing hole conduction are elements of the third group (such

as B, Al, Ga and In). Sometimes, use is also made of elements belonging to the first group (for example, Cu and Au) which exhibit the properties of both donors and acceptors.

As a rule, impurity atoms replace atoms of the host material at lattice sites, thereby producing substitution defects. An impurity atom has a different number of electrons than a host atom and

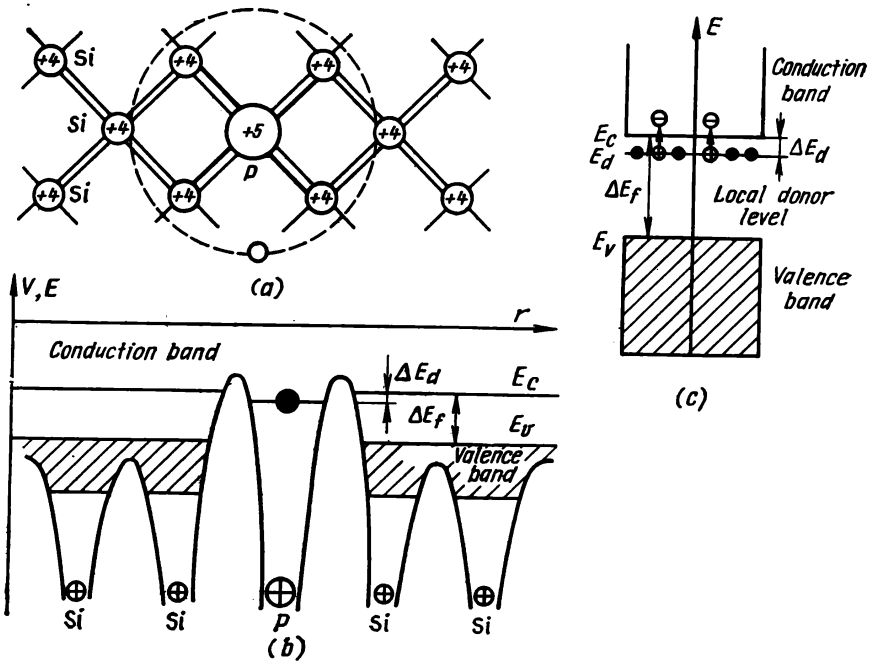


Fig. 9-3. (*a*) Impurity atom of phosphorus in a silicon crystal; (*b*) potential diar gram near the impurity atom; (*c*) energy diagram of the *N*-type semiconductor

a different atomic radius. Because of this, the periodic potential function of the crystalline structure near an impurity atom is distorted, and the atom has its energy level located within a forbidden gap (Figs. 9-3*b* and 9-4*b*). This change in the energy diagram may be explained as follows.

When a tetravalent semiconductor, say, silicon, is doped with a pentavalent element, for example, phosphorus (Fig. 9-3*a*), four valence electrons of an impurity atom form covalent bonds with the electrons of four adjacent silicon atoms. Since a covalent bond is saturated, the fifth valence electron does not take part in any of the covalent bonds; it is held to the nucleus of the phosphorus atom by the Coulomb interaction forces. Its energy state may be determined, using the hydrogen-like model of the silicon atom. We should keep in mind, however, that the fifth electron is affected not only by the Coulomb forces of the atomic radical but also by the periodic

field of the crystal structure. So, when using the hydrogen-like model, we should take into consideration the effective mass of an electron, m_n^*, and the permittivity of silicon, ε.

The solution of the Schrödinger amplitude (or, more commonly, wave) equation yields an expression [22] for the eigenvalue of E_n, the energy of an electron in an impurity atom:

$$E_n = E_c - E_{iH} z^2 m_n^* / \varepsilon^2 n^2 m \tag{9-3}$$

Here, $E_{iH} = me^4/8\varepsilon_0^2 h^2 = 13.52$ eV is the ionization energy of a hydrogen atom, E_c is the energy at the bottom of the conduction band, z is the number of unit positive charges in the atomic radical of the impurity, and $n = 1, 2, 3...$

The quantity

$$E_d = -E_{iH} z^2 m_n^* / \varepsilon^2 m \tag{9-4}$$

is the energy of the ground ($n = 1$) energy state of the donor atom as reckoned from the bottom of the conduction band. This suggests that the energy level of the donor impurity atom is in the forbidden band and its ionization energy, $\Delta E_d = E_c - E_d$, is $1/\varepsilon^2$ of the ionization energy of a hydrogen atom, E_{iH}.

For example, for silicon $\varepsilon = 12$ and $m_n^* = 0.33$, so ΔE_d is a few hundredths of an electronvolt.

The Bohr radius calculated for the "extra" electron in a pentavalent impurity is approximately hundred times r_0, the lattice constant of the host semiconductor. So, Fig. 9-3*a* only qualitatively illustrates the position of an impurity atom in the crystal structure of silicon.

When a semiconductor is doped with an acceptor impurity, one of the valence bonds with adjacent atoms remains vacant (Fig. 9-4*a*). This state cannot be regarded as a hole because the acceptor atom is electrically neutral. The ionization of the acceptor atom consists in the attraction of an electron from the valence band by the vacant valence bond

$$E_n = E_v + E_{iH} z^2 m_n^* / \varepsilon^2 n^2 m = E_v + E_a / n^2 \tag{9-5}$$

Here, E_v is the energy at the top of the valence band, and E_a is the ground state energy of the acceptor atom.

As is seen, the energy level of an acceptor impurity atom lies in the forbidden band above the top of the valence band.

The ionization energy of a trivalent acceptor atom in silicon, $\Delta E_a = E_a - E_v$, is also about several hundredths of an electronvolt.

Such impurity energy levels lying close to E_v and E_c are customarily called *shallow* levels.

If a semiconductor is doped with an impurity whose valence differs from that of the host material by more than unity, impurity atoms may be doubly or triply ionized. In this case, two or three impurity

energy levels occupy the forbidden band, and the ionization energy of the doubly or triply ionized atoms exceeds the ionization energy of a singly ionized atom, as follows from Eqs. (9-4) and (9-5) involving z^2.

The energy levels of some most common impurities are given in Table 9-2.

The population of impurities in semiconductors (N_d and N_a) is generally low, being 10^{-4} to 10^{-6}%, that is, one impurity atom

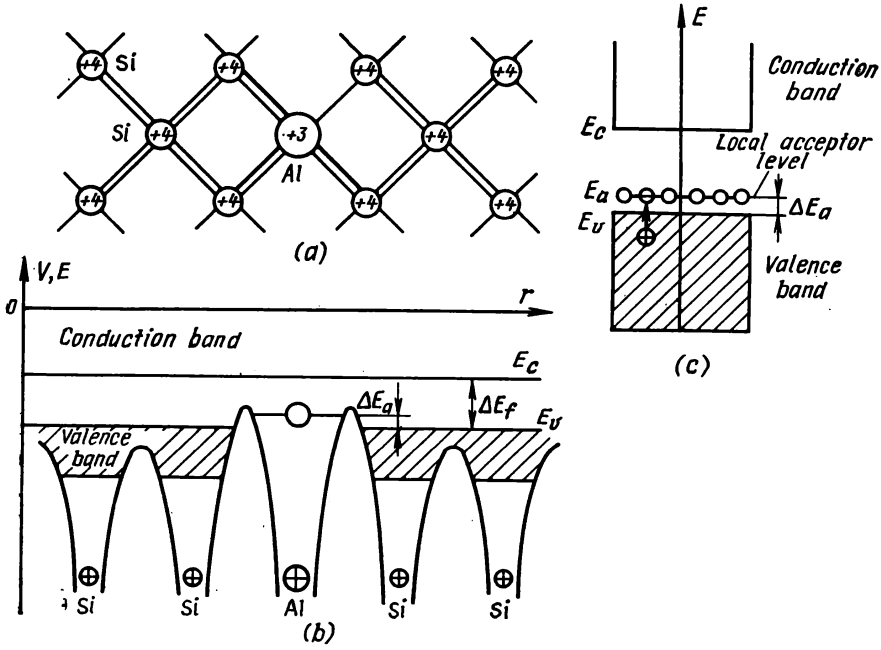


Fig. 9-4. (*a*) Impurity atom of aluminium in a silicon crystal; (*b*) potential diagram near the impurity atom; (*c*) energy diagram of the *P*-type semiconductor

per 10^6 to 10^8 atoms of the host material. This means that individual atoms of an impurity uniformly distributed in a semiconductor are spaced several hundred lattice constants r_0 apart. With such concentrations, the wave functions of adjacent impurity atoms do not practically overlap, and all impurity atoms are at a common discrete energy level. This level is often called *local* because it exists only near impurity atoms (Figs. 9-3*c* and 9-4*c*).

Impurity Energy Bands. Sometimes, the concentration of impurities in semiconductors is raised to 10^{-2} or 1%. In this case, the wave functions of adjacent impurity atoms cannot be regarded as nonoverlapping. The impurity atoms form a system of interacting particles, and the discrete energy level of the impurity becomes an *impurity energy band*. With shallow impurity levels, such a band overlaps

Table 9-2

Impurity	Group	Energy level, eV	
		Germanium	Silicon
Boron	III	$E_v+0.0104$	$E_v+0.046$
Aluminium	III	$E_v+0.0102$	$E_v+0.067$
Indium	III	$E_v+0.0112$	$E_v+0.154$
Phosphorus	V	$E_c-0.0120$	$E_c-0.046$
Arsenic	V	$E_c-0.0127$	$E_c-0.056$
Copper	I	$E_v+0.04$	$E_v+0.49$
		$E_v+0.33$	$E_v+0.24$
Gold	I	$E_v+0.05$	$E_v+0.35$
		$E_v+0.16$	$E_v+0.54$
		$E_c-0.29$	—
		$E_c-0.04$	—

the adjacent valence or conduction band. Thus, the energy-level diagram plotted for this case corresponds to that of metals (Fig. 9-5).

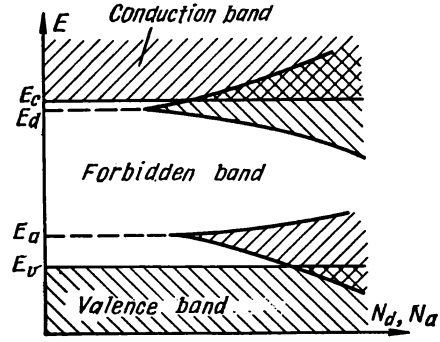


Fig. 9-5. Splitting of a local impurity level into an impurity energy band

The electrical properties of such highly alloyed semiconductors approach those of metals. That is why they are often called *semimetals.*

9-3. GENERATION AND RECOMBINATION OF FREE CHARGE CARRIERS IN SEMICONDUCTORS

For their operation, semiconductor devices depend on the motion of free electric charge carriers which are particles not bounded with separate atoms and capable of moving freely in the crystal structure. In semiconductors, free charge carriers may be formed due to either

breaking-off of valence bonds and separation of valence electrons from atoms of a semiconductor, or ionization of impurity atoms. The formation of free particles is called *generation of free charge carriers*, and the restoration of neutral atoms is known as *recombination.*

Types of Generation and Recombination of Free Carriers. Generation of free charge carriers requires that some energy be expended to "liberate" particles from their bonds. This energy may be supplied to a crystal in different forms, namely as heat (*thermal generation*), kinetic energy of moving particles (*impact generation* or *ionization*), electric field energy (*field generation* or *ionization*), photon energy of incident light (*light generation*), etc.

Recombination of particles is accompanied by dissipation of energy which may be radiated as photons (*radiative recombination*) or accepted by the crystal lattice as phonons (*nonradiative recombination*). Recombination depends on a number of factors, such as the density of free particles, the character of their motion, presence and population of impurities and some other imperfections in the crystal structure, surface conditions, and so on. It is customary to differentiate *interband* (or *band-to-band*) *recombination*, *recombination through local impurity levels* (*recombination traps*), and *surface recombination.* Now we shall discuss the main types of generation and recombination of free charge carriers in more detail.

Thermal Generation of Hole-Electron Pairs. At $T = 0$ K, a pure (intrinsic) semiconductor has no electrons in the conduction band. All valence electrons are localized near atoms and, in pairs, form covalent bonds between adjacent atoms. All of the energy levels in the valence band are occupied. Valence electrons move at random and break through the potential barriers from one potential well to another (*tunnel effect*) without changing their energy, but at any instant all energy states remain filled.

Because of thermal agitation at $T > 0$ K, any valence bond between two atoms may be broken. One of the electrons taking part in the covalent bond may acquire an energy exceeding the energy associated with that bond and become free. On a two-dimensional lattice pattern of silicon (Fig. 9-6*a*), this may conventionally be shown as a broken covalent bond (each bond is shown by two lines between atomic radicals of Si) and an electron freely moving among the vertices of the crystal lattice. The electron leaves behind an unfilled valence bond and an uncompensated charge equal to the electron charge. As is known from a course on physics, such a state is called a *hole.* In the energy-level diagram, the above process may be shown as an electron jumping from the valence band into the conduction band. In the valence band, it leaves an empty energy state thereby constituting a hole (Fig. 9-6*b*). This transfer requires some energy which at least should be equal to the width

of the forbidden band ($E \geqslant \Delta E_f$). The transfer of an electron into the conduction band necessarily produces a pair of charges, namely a negative electron in the conduction band and a positive hole in the valence band. That is why the process is called generation of hole-electron pairs. Both charges are mobile. The free electron moves at random among the vertices of the crystal lattice. In the general case, this motion is accompanied by interaction with other particles

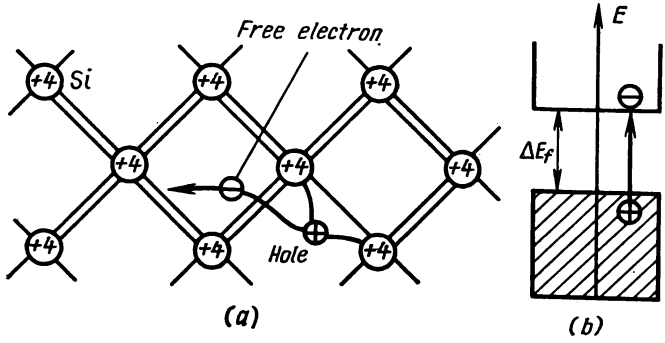


Fig. 9-6. Generation of mobile electron-hole pairs

and changes in the electron energy. So, in the energy-level diagram, it may be represented as a random motion up or down, according

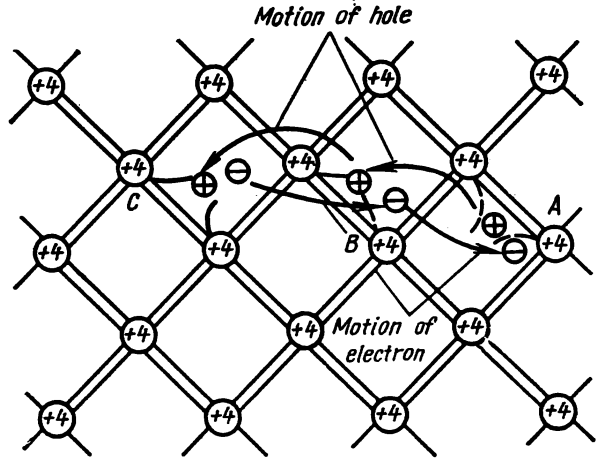


Fig. 9-7. Motion of a hole in a crystal

as the energy of an electron increases or decreases during its motion. The motion of a hole in space depends on the finite probability that the electron in a broken covalent bond may be replaced by random electrons from adjacent atoms. Referring to Fig. 9-7, the successive transition of electrons from atom C to atom B, to atom A, and so

on is equivalent to the motion of a hole in the opposite direction. In the energy-level diagram, this is represented as a successive replacement of vacant energy states in the valence band and the respective reverse motion of a hole. It should be kept in mind, however, that in a one-dimensional energy-level diagram, we can show only changes in the energy state of a particle, irrespective of whether it is present in this region of the crystal structure or not.

Thus, generation of hole-electron pairs is seen to be formation of positive and negative mobile particles which can carry electric charges.

Semiconductors not containing any impurities are called *intrinsic*, or *i-type*, *semiconductors*. Accordingly, free electrons and holes in intrinsic semiconductors are symbolized n_i and p_i. In an intrinsic semiconductor of which generation of hole-electron pairs is typical, the populations of electrons and holes are equal, $n_i = p_i$.

Thermal Ionization of Impurity Atoms. In the previous section, we have learned that the ionization energy of impurity atoms is several hundredths of an electronvolt and, as a consequence, is much smaller than the width of the forbidden band ($\Delta E_d < \Delta E_f$ and $\Delta E_a < \Delta E_f$).

Naturally, ionization of impurities starts at much lower temperatures than thermal generation of hole-electron pairs in an intrinsic semiconductor.

When a donor impurity atom in the conduction band is ionized, it gives up a free electron and turns into a *positive ion*. In contrast to the transition of a valence electron from the valence band into the conduction band which takes place in generation of hole-electron pairs, no hole is formed here because all of the valence bonds near the donor atom are filled. Thus, in contrast to a hole, a positive impurity ion is immobile and, as a consequence, only free electrons (which are negative) are generated by ionization of donors. Such semiconductors are called *N-type semiconductors*.

Due to thermal agitation of an acceptor impurity, one of the valence electrons of adjacent atoms in the host silicon may break the valence bond and occupy a vacant energy level (complete a valence bond) near the acceptor atom. As a result, a hole is constituted near the adjacent atom which had the electron in question, and the acceptor atom becomes an *immobile negative ion*. So, ionization of acceptors results only in the production of positive holes. Accordingly, such semiconductors are called *P-type semiconductors*.

Majority and Minority Carriers. Generally, extrinsic semiconductors, that is, those doped with impurities, contain both donors and acceptors. For example, an *N*-type semiconductor may have an insignificant proportion of acceptor atoms. Even at low temperatures, it has a small number of holes in addition to free electrons. Besides, in any extrinsic semiconductor, ionization of impurities

is accompanied by generation of hole-electron pairs. As a result, N-type semiconductors always contain a certain number of holes in addition to a great number of free electrons, and P-type semiconductors have an insignificant number of free electrons. Electrons in N-type semiconductors and holes in P-type semiconductors are customarily called *majority carriers*; in contrast, holes in N-type materials and electrons in P-type materials are termed *minority carriers*. Concentrations of majority carriers are symbolized n_n and p_p, and those of minority carriers, n_p and p_n.

Charge carriers are called majority carriers because their population is high and they make a major contribution to conduction.

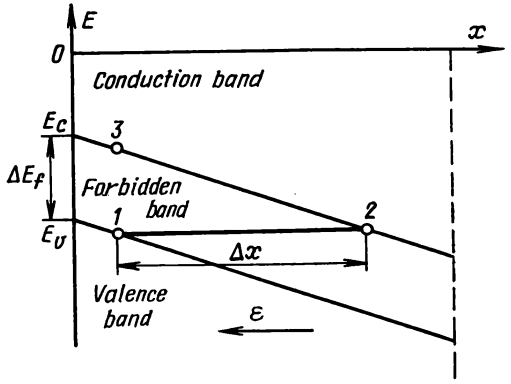


Fig. 9-8. Energy diagram of a semiconductor in a strong electric field

Impact Generation. This takes place in the presence of electric fields about 10^5 to 10^6 V/cm strong. In a uniform semiconductor, it is rather difficult to develop such a strong field by means of an external potential difference alone, because even in a crystal as small as about 0.1 mm in size, the potential difference must be several hundred or thousand volts. In semiconductor devices, however, such electric fields can exist in P-N transition or other nonuniform regions. Because of this, impact generation provides the basis for operation of some semiconductor devices.

In so strong an electric field, free particles acquire a kinetic energy sufficient to break valence bonds in atoms of a semiconductor.

Impact generation may result not only from breaking-off of valence bonds in the host semiconductor, but also from ionization of impurity atoms. In this case, a lower energy is required, so impact ionization of impurity atoms takes place in electric fields with a strength of $\mathscr{E} \geqslant 10$ V/cm.

Field Generation. This also takes place in very strong electric fields, but it differs from impact generation. As is known from a course on physics, an applied electric field causes the boundaries of energy bands in a semiconductor to shift. Figure 9-8 shows the energy-level diagram of a part of a semiconductor crystal placed in a strong electric field. For simplicity, we assume that the boundaries of the energy bands are shifted linearly. As a result of this shift, the energy of the electron at point *1* in the top of the valence band is equal to the energy of the electron at point *2* in the bottom

of the conduction band. The two energy states are separated by a triangular energy barrier *1-2-3* which is ΔE_f high and Δx wide. If the electric field strength $\mathscr{E}$ is sufficiently high (about 10^6 V/cm), then the width Δx may be rather small and comparable with the wavelength of the electron wave function at point *1*. Such fields may be developed in transition regions. In this case, the probability of a tunnel transition of a particle is finite, provided energy level *2* is not occupied, which is quite possible with a low concentration of electrons in the conduction band.

As we know from a course on physics, the probability that a particle will move over a narrow energy barrier is defined by

$$P = (\text{approx.}) \exp\left(-\frac{4\pi}{h}\sqrt{2m^*\Delta E}\,\Delta x\right) \qquad (9\text{-}6)$$

where ΔE is the height of the energy barrier, Δx is the width of the barrier, and m^* is the effective mass of the particle.

Referring to Fig. 9-8, ΔE is equal to ΔE_f, and

$$\Delta x = \Delta E_f / e\mathscr{E} \qquad (9\text{-}7)$$

Substituting these quantities into Eq. (9-6), we get

$$P = (\text{approx.}) \exp\left(-\frac{4\pi}{eh\mathscr{E}}\sqrt{2m^*(\Delta E_f)^3}\right) \qquad (9\text{-}8)$$

This suggests that the probability of a tunnel transition increases with increasing field strength $\mathscr{E}$ and is higher for semiconductors having a narrow forbidden band.

In real semiconductor devices, the tunnel effect in strong electric fields may cause considerable generation of free carriers.

Light Generation. This occurs when a semiconductor is exposed to light. Two types of light generation are differentiated, namely *bipolar* and *monopolar*. In the former, the energy of a photon absorbed by a semiconductor, $h\nu$, is expended to break the valence bonds in the atoms of the semiconductor. As this takes place, both electrons and holes are formed, which is why this type of light generation is called bipolar.

A valence bond may be broken only if the energy of the absorbed photon is $h\nu \geqslant \Delta E_f$. Thus, for any semiconductor, bipolar generation occurs only in a certain portion of the light spectrum. This portion is called the *absorption spectrum*.

In monopolar generation, absorption of a photon is accompanied by ionization of impurity atoms. Here, the energy of an absorbed photon should be $h\nu \geqslant \Delta E_d$ or $h\nu \geqslant \Delta E_a$. At room temperature, shallow-level impurities are ionized almost completely. So, monopolar generation in the case of extrinsic semiconductors is important only with relatively deep impurity levels.

In more detail, absorption of light by semiconductors and accompanying generation of free charge carriers are discussed in Chapter 14.

State of Equilibrium. Generation and recombination of free charge carriers take place simultaneously. When a semiconductor is held at a constant temperature, these processes are in dynamic equilibrium, that is, equal numbers of like charge carriers appear and vanish per unit time. Otherwise, the number of electrons and holes would increase or decrease with time.

If we symbolize the number of electron-hole pairs generated per unit time (the *rate of generation*) g_0, and the number of pairs eliminated over the same time and in the same volume (the *rate of recombination*) r_0, then we may define the condition for this dynamic equilibrium as follows

$$r_0 = g_0 \tag{9-9}$$

Here, the concentrations of electrons and holes should naturally remain constant. Such charge carriers are in thermal equilibrium with the crystal lattice, so they are customarily called *equilibrium carriers*, and the equilibrium concentrations of electrons and holes are symbolized n_0 and p_0, respectively. Under these conditions, the system itself is said to be in equilibrium.

Accordingly, the number of electron-hole pairs recombining per unit time is proportional to their number densities

$$r_0 = \gamma_r n_0 p_0 \tag{9-10}$$

where γ_r is the *recombination factor*.

Recalling Eq. (9-10), we may re-write Eq. (9-9) defining the condition for dynamic equilibrium as

$$\gamma_r n_0 p_0 = g_0 \tag{9-11}$$

Lifetime of Free Charge Carriers. After a free particle has been generated, it takes some time to recombine with another particle of the opposite sign. This time depends on the probability of collision with a particle of the opposite sign, the probability of dissipation of the energy released due to recombination, etc. The mean life of a particle in a free state is called its *lifetime*.

As can be seen from Eq. (9-10), the quantity $\gamma_r p_0$ is the probability of electron recombination per unit time, and $\gamma_r n_0$ is the probability of hole recombination per unit time. The lifetime of a particle may be defined as the reciprocal of its recombination probability. Thus, the lifetime of an electron is

$$\tau_n = 1/\gamma_r p_0 \tag{9-12}$$

and the lifetime of a hole

$$\tau_p = 1/\gamma_r n_0 \tag{9-13}$$

Recalling Eq. (9-10), we may re-write the above expressions as

$$\tau_n = n_0/r_0 \tag{9-14}$$

$$\tau_p = p_0/r_0 \tag{9-15}$$

Using the concept of the lifetime of a particle, we may re-write Eq. (9-9) defining the condition for dynamic equilibrium as

$$n_0/\tau_n = p_0/\tau_p = g_0 \tag{9-16}$$

State of Non-Equilibrium. Some external factors (incident light, rapid heating, bombardment by particles, etc.) may cause intensive generation of free charge carriers in a certain crystal volume. In this case, the steady state will be upset, that is, the number densities of particles will increase or decrease, thereby *upsetting the dynamic equilibrium*

$$n = n_0 + \Delta n \tag{9-17}$$

$$p = p_0 + \Delta p \tag{9-18}$$

where Δn and Δp are the so-called *excess free electrons* and *holes.*

Any disturbed system seeks to attain a new dynamic equilibrium (in the case of a sustained disturbance) or restore the original dynamic equilibrium (after the disturbance has ceased). In our case, this happens owing to the exchange of energy between electrons and holes and the crystal lattice, an increased recombination rate in the volume with excess particles, and also their diffusion from this volume due to the concentration gradient.

The free charge carriers formed because of disturbances may have an energy considerably exceeding the energy of equilibrium particles (about kT). This means that, say, electrons may occupy higher energy levels in the conduction band, which they usually do at a higher temperature. Therefore, such electrons are called "hot". As they move about at random, free electrons interact with the crystal lattice and dissipate their energy.

Diffusion of Non-Equilibrium Carriers. Their Lifetime. An increase and a decrease in the number of charge carriers in a volume with non-equilibrium particles due to their recombination may be described as

$$dn/dt = g_n - \gamma_r n p_0 \tag{9-19}$$

$$dp/dt = g_p - \gamma_r p n_0 \tag{9-20}$$

where g_n and g_p are the recombination rates of electrons and holes, respectively.

After external disturbances have ceased to affect the system, it tends to equilibrium: the equilibrium carriers diffuse and gradually recombine.

Assuming that $g_n = g_p = g_0$ and recalling Eq. (9-11), we may re-write Eq. (9-19) as follows

$$-dn/dt = \gamma_r (np - n_0 p_0) \tag{9-21}$$

Substituting Eqs. (9-17) and (9-18) into the above expression, we get

$$-dn/dt = \gamma_r (n_0 \Delta p + p_0 \Delta n + \Delta n \Delta p) \quad (9\text{-}22)$$

Most often, $\Delta n < n_0$ and $\Delta p < p_0$, so the third term in the brackets in Eq. (9-22) may be neglected. Also, if charges of both signs are generated at the same rate ($g_p = g_n$), then $\Delta p = \Delta n$ and Eq. (9-22) takes the form of

$$-dn/dt = (\gamma_r n_0 + \gamma_r p_0)(n - n_0) \quad (9\text{-}23)$$

Recalling Eqs. (9-12) and (9-13), we get

$$-dn/dt = (1/\tau_n + 1/\tau_p)(n - n_0) \quad (9\text{-}24)$$

The quantity

$$1/(1/\tau_n + 1/\tau_p) = \tau \quad (9\text{-}25)$$

is called the *lifetime of non-equilibrium carriers.*

Recalling Eqs. (9-17) and (9-25), we re-write Eq. (9-24) as

$$dn/dt = -(n - n_0)/\tau = -\Delta n/\tau \quad (9\text{-}26)$$

The solution of the above equation gives

$$\Delta n = \Delta n\,(0) \exp(-t/\tau) \quad (9\text{-}27)$$

where $\Delta n\,(0)$ is the non-equilibrium density at the instant when disturbance ceases. Thus, the number of non-equilibrium electrons decreases to 1/e during time τ. A similar conclusion may be drawn for holes by solving Eq. (9-20).

Interband Recombination. This occurs when an electron jumps from free into a bounded (valence) state. In the energy-level diagram, this may be shown as the transition of an electron from the conduction into the valence band (transition *1* in Fig. 9-9*a*). With such a transition, the electron naturally releases some energy equal to the width of the forbidden band. In the case of interband recombination, energy may be dissipated as photons (*radiative recombination*) or as phonons (*nonradiative recombination*). In many respects, the character of interband recombination depends on the structure of the energy bands in the semiconductor. Radiative interband recombination is typical of those semiconductors whose $E\,(k)$ function has its extrema at the same value of the wave number k and, as a consequence, of the quasimomentum $p = (h/2\pi)\,k$, for example, InSb (Fig. 9-9*b*). When an electron moves from the conduction band to the valence band, its pulse value does not change (*direct transition*); the energy equal to ΔE_f is most often released as a photon.

When the extrema of the $E\,(k)$ function correspond to different values of the quasimomentum as, say, in silicon (Fig. 9-9*c*), we speak mostly of nonradiative recombination (*indirect transition*).

released in an indirect transition should absorb not only the energy equal to ΔE_f but also the excess (or deficiency) value of the quasi-momentum. A photon having a very small momentum is unable to absorb the excess or deficiency in the momentum of an electron, as it is by several orders of magnitude higher. This excess momentum is usually absorbed by a phonon whose quasimomentum may reach a considerable value. The excess energy may be absorbed by a photon or, under certain conditions, a phonon.

The latter is characteristic of most semiconductors. The probability of nonradiative interband recombination is insignificant because

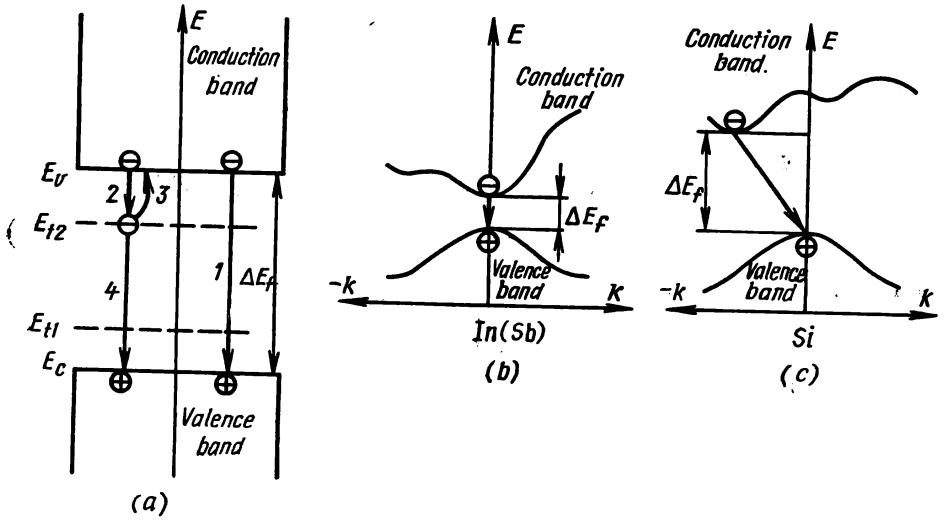


Fig. 9-9. (*a*) Trap recombination of particles, (*b*) and (*c*) energy bands of InSb and Si

the energy ΔE_f is rather significant (being 1 to 3 eV) as compared with kT and, therefore, cannot be absorbed by the crystal lattice always.

Trap Recombination. This type of recombination is most likely to occur because of irregularities in the crystal structure known as *impurity centres*. As has already been noted in Sec. 9-2, impurities generally used in alloyed semiconductors (such as B, Al, P and As) form shallow local levels ($\Delta E_d \approx \Delta E_a \approx 0.01$ eV). Atoms of these impurities cannot markedly affect recombination because the energy released when an electron moves from the conduction band to the acceptor level is almost equal to ΔE_f. However, such impurities as copper and gold form a number of impurity levels some of which are located near the middle of the forbidden band (see Table 9-2). For example, a monovalent atom of gold replacing a germanium atom in the crystal structure produces three local levels two of which ($E_v + 0.16$ and $E_c - 0.29$) are sufficiently far away from

the bottom of the conduction band and the top of the valence band. Generally, this atom may capture three electrons which it needs to fill the valence bonds. The probability of a transition from the conduction band to the local level E_{t_2} (transition *2* in Fig. 9-9*a*) is higher than that of an interband transition. Later, this electron may return to the conduction band (transition *3*) or move to the valence band (transition *4*) which is equivalent to the transition of a hole from the valence band to the impurity level. In the latter case, recombination takes place step by step, and the energy ΔE_f is released in portions.

Impurity centres which can capture mobile charge carriers and thus contribute to their recombination are called *recombination traps.*

Recombination traps may be constituted not only by impurity atoms but also by other imperfections in the crystal structure.

Surface Recombination. This is a variety of trap recombination. Because of irregularities in the crystal structure, so called *surface states with energy levels* in the forbidden band appear at the surface of the crystal. In addition, surface states are formed due to oxidation of the surface and absorption of foreign atoms. These atoms may be donors and acceptors. As they capture or give up electrons, the atoms compose a negative or a positive layer at the surface. This gives rise to an electric layer near the surface with a surface potential of V_s, and the band boundaries are bent near the surface. The energy levels of surface states whose density is as high as 10^{10} to 10^{12} cm^{-2} may serve as recombination traps. They affect the lifetime of carriers at both the surface and some depth inside the crystal.

Charge Neutrality. In general, any crystal is electrically neutral, irrespective of the character and rate of generation and recombination of mobile equilibrium or non-equilibrium charge carriers. The total charge of mobile and immobile negative particles is always compensated for by the total charge of mobile and immobile positive particles. When the unlike charges are equal, a semiconductor is said to be *electrically neutral.*

For example, in an intrinsic semiconductor, the total charge of free electrons is compensated for by the total charge of holes

$$n_i = p_i \tag{9-28}$$

In an N-type extrinsic semiconductor, the charge of free electrons is neutralized by the total positive charge composed of the hole charges and the charges of ionized donor atoms

$$n_n = p_n + N_d^+ \tag{9-29}$$

At operating temperatures, when generation of electron-hole pairs is negligible, the hole charge may be ignored

$$n_n = (\text{approx.})\, N_d^+ \tag{9-30}$$

Similarly, for a P-type extrinsic semiconductor we may write

$$p_p = n_p + N_a^- \tag{9-31}$$

whence, for operating temperatures

$$p_p = (\text{approx.})\, N_a^- \tag{9-32}$$

For an extrinsic semiconductor containing both donor and acceptor impurities at $N_d > N_a$

$$n_n + N_a^- = p_n + N_d^+ \tag{9-33}$$

and at $N_a > N_d$

$$p_p + N_d^+ = n_p + N_a^- \tag{9-34}$$

9-4. PARTICLE STATISTICS AS APPLIED TO SEMICONDUCTORS

The band theory of solids describes their energy spectrum as a series of energy bands each of which contains energy states allowed (or permitted) for particles. The number of discrete energy states within a band is equal to the number of atoms making up a crystal lattice, and the band width, or the energy gap between its highest and lowest levels (the top and bottom of a band) depends on how much the wave functions of adjacent atoms overlap. A good many crucial problems, such as the manner in which the energy states are disposed in a band and the distribution of particles among these states, can only be treated by the methods of statistical physics. As applied to an ensemble of particles, the principal goal of statistics is to determine the probability that an ith particle has an energy in the interval between E and $E + \Delta E$.

Classical Statistics. Statistical methods were first applied to the ideal gas by Boltzmann and elaborated by Maxwell (the classical Maxwell-Boltzmann statistics).

In accordance with classical statistics, the energy distribution function for N particles has the form

$$\Phi(E) = dN/dE = C\sqrt{E}\exp(-E/kT) \tag{9-35}$$

Here,

$$C = 2n/\pi^{1/2}(kT)^{3/2} \tag{9-36}$$

where n is the number of particles per unit volume (number density of particles).

In a course on physics, the chemical potential, χ, is defined as the amount by which the free energy of a system varies as the number of particles in it is changed by one, on the condition that the system volume V and the system temperature T remain unchanged. Thus, at V = constant and T = constant, the free energy change is a univalued function of the number density (concentration) of particles

in a system. When the particles are uniformly distributed throughout a system, the chemical potential is the same for all of the system, that is, χ=constant and grad χ=0. If, on the other hand, the number density of particles is different in various parts of a system, the chemical potential will likewise be different there, and grad $\chi \neq 0$. As will be recalled, the difference in number density of particles between various regions in a system gives rise to their diffusion, that is, migration from a high-concentration to a low-concentration region.

Classical statistics had been developed before the advent of quantum mechanics. Quite obivously, it could not take into account the quantum-mechanical properties of particles and particle interaction. For example, classical statistics overlooked Pauli's exclusion principle, by allowing any number of particles to occupy any energy state. That was the reason why classical statistics was unable to explain plausibly the physical processes occurring in such ensembles of particles where quantum-mechanical properties play a leading part, such as the electron gas of metals.

Quantum Particle Statistics. Initially, this statistics was formulated by Bose and Einstein for ensembles of integral-spin particles (for example, photons) called *bosons*. For ensembles of halfinteger-spin particles (including electrons) which obey Pauli's exclusion principle and are called *fermions*, the statistics was formulated by Fermi and Dirac.

The Fermi-Dirac statistics may in short be stated as follows. The particles forming a system are indistinguishable: an exchange of energy states between any two electrons will leave the energy of the system undisturbed. The energy spectrum is discrete; two adjacent energy states will always differ by a finite amount, however small. Any energy state may be occupied only by one particle or by two particles provided their spins are opposite.

The Energy State Density Function. In order to find the probability that a particle has an energy in the interval from E to $E + \Delta E$, one above all needs to know the energy state density law. In other words, one needs to know the manner in which the number Z of energy states varies within an unvarying energy interval, ΔE, according to the position of this interval on the energy axis. The solution of this problem reduces to the following expression for the energy state density [22]

$$S_q(E) = dZ/dE = A\sqrt{|E - E_0|} \tag{9-37}$$

where

$$A = (4\pi/h^3)(2m^*)^{3/2} \tag{9-38}$$

and E_0 is the energy at the origin. For example, for the conduction band we assume the origin to be the energy corresponding to the

bottom of the band, $E_c = 0$, the function $S_q(E)$ will have the form shown in Fig. 9-10. For a band other than conduction, for example, the valence band, the state density increases in proportion to $\sqrt{|E - E_{01}|}$, from the bottom of the band upwards, while near the top of the band it decreases in proportion to $\sqrt{|E - E_{02}|}$.

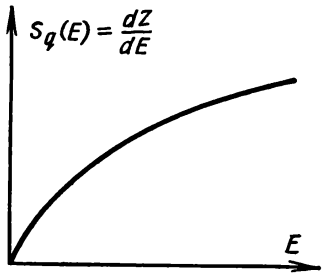


Fig. 9-10. Energy state density function in quantum statistics for an open band

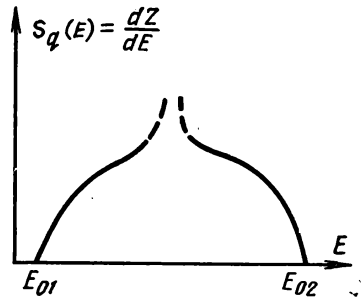


Fig. 9-11. Energy state density function in quantum statistics for a closed band

In these expression, E_{01} and E_{02} are the energies corresponding to the top and bottom of a band, respectively (Fig. 9-11).

The Fermi-Dirac Distribution Function. As will be recalled from a course in physics, this function is defined as

$$F_q(E) = dN/dZ = 1/\{\exp[(E - E_F)/kT] + 1\} \qquad (9\text{-}39)$$

where E_F is a certain fixed energy level to be defined later.

Let us examine the above function on assuming that the temperature of the system tends to zero, $T \to 0$:

— if $E < E_F$, then $\exp(E - E_F)/kT \to \exp(-\infty) \to 0$ and $F_q(E) = 1$;

— if $E > E_F$, then $\exp(E - E_F)/kT \to \exp(\infty) \to \infty$ and $F_q(E) = 0$;

— if $E = E_F$, then $\exp(E - E_F)/kT \to \exp(0) \to 1$ and $F_q(E) = 1/2$. A plot of Eq. (9-39) for $T = 0$ K is shown in Fig. 9-12.

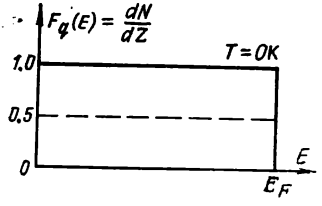


Fig. 9-12. Fermi-Dirac distribution function in quantum statistics

Fermi Level. Now we shall explain the physical meaning of the function $F_q(E)$. When offered the choice of Z likely states, particles will tend to occupy states with the least energy. These occupied states will extend from zero upwards to E_F. In accordance with the Pauli exclusion principle, if two particles occupy the same energy state, they must differ in spin. As a consequence, $Z = N/2$ states will be occupied by particles in the energy interval from zero to E_F; states with energy $E > E_F$ will be left unoccupied. The energy E_F is, therefore, the highest energy that a particle can have in a system

at $T = 0$ K. It is called the *Fermi energy* or the *Fermi level.* It will be more correct to interpret the function $F_q(E)$ statistically. From this viewpoint, at $T = 0$ K the probability that particles will occupy states with energy $E < E_F$ is unity, and the probability that they will occupy states with energy $E > E_F$ is zero.

As is seen from Fig. 9-12, the Fermi level at $T = 0$ K separates the occupied from unoccupied states. That is, the Fermi level separates the energy states for which the occupancy probability is more than 1/2, from the states for which the occupancy probability is less than 1/2. This statement holds for any system temperature.

The quantity E_F/e is often called the *electrochemical potential.* The chemical potential, χ, that is, the quantity uniquely determining the concentration of particles in a system of electrically neutral particles was defined earlier. In an ensemble of charged particles (on the condition that T = constant and V = constant), a change in the number of particles by one gives rise not only to the concentration gradient but also to the potential gradient (grad $\varphi \neq 0$). The disturbed equilibrium results in both diffusion of particles in a system and their drift, that is, their motion under the action of an electric field.

As is seen, the condition for equilibrium of a system of charged particles is that the chemical potential, χ, and the electrochemical potential, φ, must be constant.

The Fermi energy (the electrochemical potential) is a quantity combining the two conditions. In a system of charged particles, the condition E_F = constant for the entire system uniquely defines the condition for equilibrium of the system, although it is not necessary that χ and φ should be constant. In other words, a system at equilibrium (E_F = constant) may have both a concentration and a potential gradient and, as a consequence, show diffusion and ohmic (drift) currents of particles, but these two currents will neutralize each other, so that the resultant current will be zero.

The Energy Distribution Function. This function may be obtained by multiplying the functions $S_q(E)$ and $F_q(E)$

$$\Phi_q(E) = S_q(E)\, F_q(E) = (dZ/dE)(dN/dZ) = dN/dE \qquad (9\text{-}40)$$

Substituting Eqs. (9-39) and (9-37) into Eq. (9-40), we get

$$\Phi_q(E) = A\left(\sqrt{|E - E_0|}\right)/\{\exp[(E - E_F)/kT] + 1\} \qquad (9\text{-}41)$$

For an open band, the function $\Phi_q(E)$ at $T = 0$ K has the form of Fig. 9-13*a*. The energy states are filled up to E_F. A majority of particles have energies close to E_F. Particles with low energies are few because at a low E the energy state density is small. If all of the energy states are filled, the function $\Phi_q(E)$ for a closed band (Fig. 9-13*b*) follows the function $S_q(E)$ shown in Fig. 9-11. If the

number of particles is smaller than the number of possible states ($E_F < E_0$), then the closed band is filled only partially (Fig. 9-13*c*).

Change in the Distribution Functions at T > 0 K. Let us assume that a system of particles is held at a nonzero temperature ($T_1 > 0$ K). As is seen from Eqs. (9-37) and (9-38), the function $S_q(E)$ does not depend on temperature. Two other functions, $F_q(E)$ and $\Phi_q(E)$, should change with rising temperature. As the temperature increases ($T_1 > 0$ K), particles receiving additional energy should jump into higher energy states. Such transitions, however, can be carried out only by those particles for which the new higher energy states are vacant. In other words, the additional thermal energy

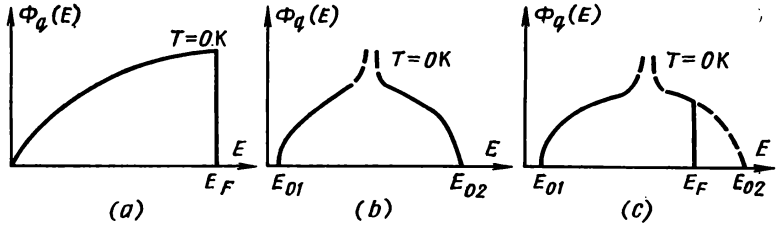


Fig. 9-13. Energy distribution function in quantum statistics
(*a*) open band; (*b*) closed filled band; (*c*) closed unfilled band

will only be received by those particles whose energy at $T = 0$ K is kT_1 below the energy E_F. Particles which at $T = 0$ K are at low energy states cannot carry out such transitions and, as a consequence, receive additional thermal energy because the states which they should fill are occupied by other particles. Thus, a change in the function $F_q(E)$ caused by rising temperature is to be expected near the energy level $E = E_F$. To the left of E_F, the probability that particles will occupy energy states becomes smaller than unity, and to the right of E_F, it turns out to be non-zero (Fig. 9-14*a*). For example, when $E = E_F - 4kT$, $F_q(E) =$ (approx.) 0.98, and, at $E = E_F + 4kT$, $F_q(E) =$ (approx.) 0.02. It is obvious that, as the temperature increases, the function varies within a wider interval. However, the probability that the Fermi level, E_F, will be occupied remains unchanged and equal to 1/2 at any temperature.

A change in the function $F_q(E)$ brought about by increasing temperature is accompanied by a change in the function $\Phi_q(E)$ (Fig. 9-14*b* through *d*). For a completely filled closed band, the function $\Phi_q(E)$ can vary only if the thermal energy is enough for particles to move to vacant levels in the adjacent allowed energy band (Fig. 9-14*d*).

Degenerate and Non-Degenerate Electron Gases. Let us assume that in Eq. (9-41)

$$\exp[(E - E_F)/kT] \gg 1 \qquad (9\text{-}42)$$

Then, the unity in the denominator in Eq. (9-41) may be neglected and, assuming that $E_0 = 0$, the expression reduces to

$$\Phi_q(E) = (\text{approx.})\, A \sqrt{E} \exp(E_F/kT) \exp(-E/kT) \qquad (9\text{-}43)$$

Equation (9-42) holds if, in particular,

$$\exp(E_F/kT) \ll 1 \qquad (9\text{-}44)$$

Let us define this exponential cofactor by taking an integral of the function $\Phi_q(E)$, which is equal to the concentration of particles

$$n = \int_0^\infty \Phi_q(E)\, dE \qquad (9\text{-}45)$$

Hence, for the exponential cofactor we get

$$\exp(E_F/kT) = nh^3/2\,(2\pi mkT)^{3/2} \qquad (9\text{-}46)$$

Substituting this expression into Eq. (9-43) and recalling Eq. (9-38) we find

$$\Phi_q(E) = 2n \sqrt{E} \exp(-E/kT)/\pi^{1/2}\,(kT)^{3/2} \qquad (9\text{-}47)$$

which is identical to the energy distribution function (9-35) defined by the classical Maxwell-Boltzmann statistics. Thus, if the condi-

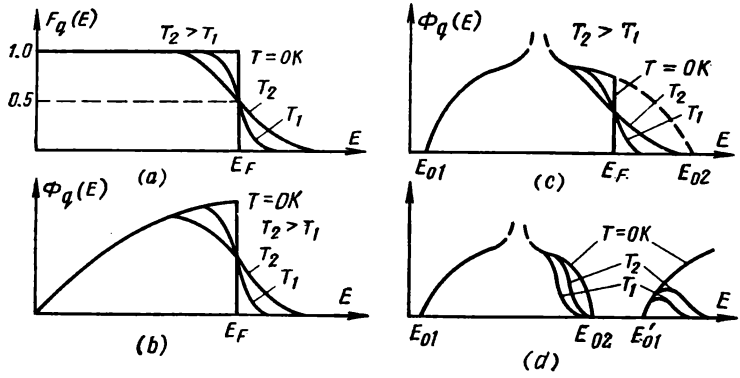


Fig. 9-14. Variations in distribution functions with temperature

tion of Eq. (9-42) is satisfied, the quantum statistics distribution function transforms to the respective function of classical statistics. The ensemble of electrons for which inequality (9-42) holds and, as a consequence, which can be described in terms of classical statistics is called a *non-degenerate electron gas.* On the other hand, the electron gas for which inequality (9-42) does not hold is called a *degenerate gas.*

As is seen from Eq. (9-46), an electron gas may become degenerate either at a low temperature or at a high concentration of particles. The degeneracy criterion may be given by the concentration of particles

$$n_{deg} = 3.67 \times 10^{15} T^{3/2} \tag{9-48}$$

In semiconductor devices, inequality (9-42) holds in most practical cases for ensembles of free charge carriers, namely electrons in the conduction band and holes in the valence band. The point is that at temperatures close to room temperature ($T \approx 300$ K), $kT \approx$ ≈ 0.03 eV. The energy difference $E - E_F$ is equal to units or tenths of an electronvolt, because E should be assumed to take values close to the energy at the bottom of the conduction band, E_c, or the energy at the top of the valence band, E_v. As will be shown in Sec. 9-5, in the energy-level diagram of a semiconductor, the Fermi level is located in the forbidden band whose width, $\Delta E_f = E_c - E_v$, ranges between 1 and 3 eV. Thus, generally $E - E_F > kT$ and, as a consequence, $\exp [(E - E_F)/kT] \gg 1$.

9-5. CONCENTRATION OF FREE CHARGE CARRIERS IN SEMICONDUCTORS

As with Eq. (9-45), the equilibrium concentration of particles in a semiconductor may be defined as an integral of the function $\Phi_q (E) = S_q (E) F_q (E)$. To find the equilibrium concentration of electrons in the condcution band, n_0, we should integrate the function going from the energy E_c corresponding to the bottom of the conduction band.

Using Eqs. (9-37) and (9-39), we get

$$n_0 = \int_{E_c}^{\infty} 2\pi (2m_n^*/h^2)^{3/2} (E - E_c)^{1/2} dE/\{\exp [E - E_F)/kT] + 1\} \tag{9-49}$$

In the general case, this integral has no solution. In most important practical cases, the electron gas in a semiconductor is non-degenerate and, consequently, the unity in the denominator may be neglected. Now, the solution of Eq. (9-49) is

$$n_0 = N_c \exp [(E_F - E_c)/kT] \tag{9-50}$$

where

$$N_c = 2 (2\pi m_n^* kT)^{3/2}/h^3 \tag{9-51}$$

is the *effective number of states in the conduction band.*

Assuming that $E_F \to E_c$, N_c defines the maximum possible concentration of electrons in a non-degenerate semiconductor. Since n_0/N_c is the ratio of the equilibrium electron concentration to the

maximum possible concentration, the exponential multiplier in Eq. (9-50) has the sense of the probability that free electrons will occupy the conduction band.

The equilibrium concentration of holes in the valence band may be determined in a similar manner. Here, the function $F_q(E)$ in Eq. (9-49) should be replaced by the expression $1 - F_q(E)$ which is equal to $\exp[-(E_F - E)/kT]$ for a non-degenerate semiconductor.

In this case, Eq. (9-49) is solved as follows

$$p_0 = N_v \exp[(E_v - E_F)/kT] \tag{9-52}$$

where

$$N_v = 2(2\pi m_p^* kT)^{3/2}/h^3 \tag{9-53}$$

is the *effective number of states in the valence band.*

As in Eq. (9-50), the exponential multiplier in Eq. (9-52) has the sense of the probability that holes will occupy the valence band, and N_v defines the maximum possible concentration of holes in a non-degenerate semiconductor, provided $E_F \to E_v$.

Concentration of Charge Carriers in an Intrinsic Semiconductor. Equations (9-50) and (9-52) defining concentrations of electrons and holes in a non-degenerate semiconductor hold for an intrinsic semiconductor in which the number of holes is equal to that of electrons

$$n_0 = p_0 = n_i \tag{9-54}$$

Hence

$$n_i^2 = n_0 p_0 \tag{9-55}$$

Using Eqs. (9-50) and (9-52), we get

$$\begin{aligned} n_i &= \sqrt{n_0 p_0} = \sqrt{N_c N_v} \exp[-(E_c - E_v)/2kT] \\ &= 2(2\pi kT)^{3/2} (m_n^* m_p^*)^{3/4} \exp(-\Delta E_f/2kT)/h^3 \end{aligned} \tag{9-56}$$

As is seen, the concentration of carriers in an intrinsic semiconductor significantly depends on temperature (the exponential multiplier being of especial importance) and also the width of the forbidden band. For example, in silicon whose energy gap ($\Delta E_f \approx 1.1$ eV) is about half as much again as that of germanium ($\Delta E_f \approx 0.7$ eV), the intrinsic concentration at room temperature ($T \approx 300$ K) is three orders of magnitude below that in germanium: n_i (Si) = = (approx.) 2×10^{10} cm^{-3} and n_i (Ge) = (approx.) 2×10^{13} cm^{-3}.

An increase in the concentration of free carriers with rising temperature is mainly defined by the exponential member in Eq. (9-56), so, in the $\log_e n_i - 1/T$ coordinates, the function $n_i = f(T)$ is a straight line making an angle α with the x-axis, whose slope is proportional to the width of the forbidden band ΔE_F (Fig. 9-15). It is very convenient to plot the concentration of particles as a function

of temperature in semi-logarithmic coordinates because a rise of several degrees in temperature causes the concentration of charge carriers to change by several orders of magnitude.

The Fermi Level in an Intrinsic Semiconductor. Let us write condition (9-54) for an intrinsic semiconductor using Eqs. (9-50) and (9-52)

$$N_c \exp [(E_F - E_c)/kT] = N_v \exp [(E_v - E_F)/kT] \quad (9\text{-}57)$$

whence

$$E_{Fi} = (E_c + E_v)/2 + (kT/2) \log_e (N_v/N_c) \quad (9\text{-}58)$$

Substituting Eqs. (9-51) and (9-53) in the above equation gives

$$E_{Fi} = (E_c + E_v)/2 + (3kT/4) \log_e (m_p^*/m_n^*) \quad (9\text{-}59)$$

As is seen, at $T = 0$ K, the Fermi level in a non-degenerate intrinsic semiconductor is located at the centre of the forbidden band. This also holds for any temperatures because, for most practical cases, the second member is rather small ($m_n^* \approx m_p^*$).

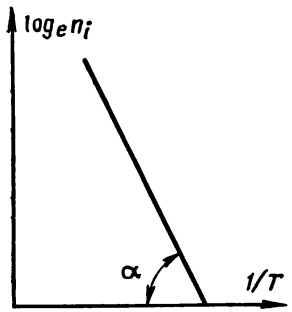


Fig. 9-15. Particle concentration in an intrinsic semiconductor as a function of temperature

The location of the Fermi level in the energy-level diagram at $m_n^* = m_p^*$ and the distribution functions are shown in Fig. (9-16). The shaded area in the upper part of the Φ_q (E) plot is proportional to the number of free electrons, n_0, and the unshaded area in the lower part of the same plot, to the number of holes, p_0, in the valence band.

Concentration of Charge Carriers in Extrinsic Semiconductors. Suppose that the concentration of pentavalent donor atoms in an N-type extrinsic semiconductor is N_d. In the general case at $T > 0$ K, a proportion of the donor atoms may be ionized. Let us designate the concentration of ionized impurity atoms as N_d^+. Then the concentration of electrons in impurity levels is $n_d = N_d - N_d^+$.

The occupancy of donor levels by electrons may be defined using an equation given in reference [22] and the function F_q (E)

$$n_d/N_d = 1/\{(1/2) \exp [(E_d - E_F)/kT] + 1\} \quad (9\text{-}60)$$

Here, E_d is the local energy level of donors.

The concentration of electrons in the conduction band, n_0, equal to the concentration of ionized donors, is defined as

$$n_0 = N_d^+ = N_d/\{2 \exp [(E_F - E_d/kT] + 1\} \quad (9\text{-}61)$$

Solving this equation simultaneously with Eq. (9-50), we get

$$n_0 = 2N_d/\{\sqrt{(8N_d/N_c)\exp[(E_c - E_d)/kT] + 1} + 1\}$$
$$= 2N_d/(\sqrt{1+D} + 1) \qquad (9\text{-}62)$$

The concentration of holes in the valence band of a P-type semiconductor may be defined in a similar way

$$p_0 = 2N_a/\{\sqrt{(8N_a/N_v)\exp[(E_a - E_v)/kT} + 1 + 1\} \qquad (9\text{-}63)$$

Let us discuss the relation between the concentration of electrons in the conduction band of a P-type semiconductor, $n_n = n_0$, and temperature, represented in Fig. 9-17 by the curve $ABCD$.

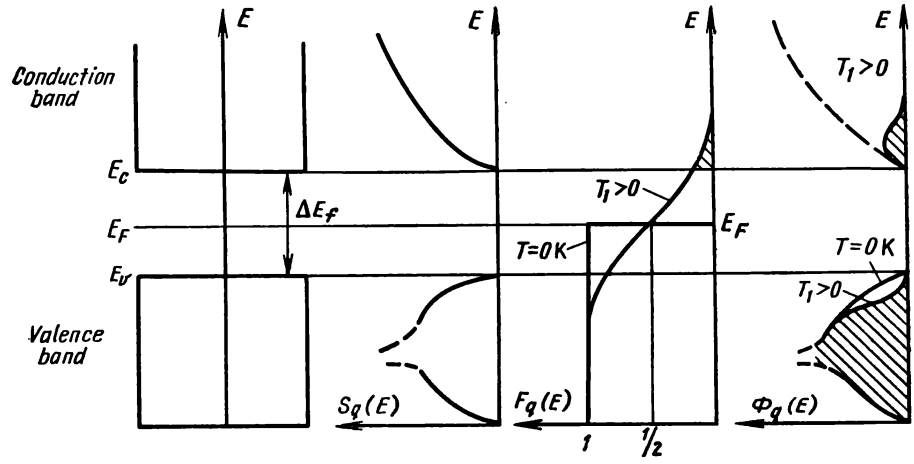


Fig. 9-16. Location of the Fermi level and variations in the distribution functions with temperature in an intrinsic semiconductor

At $T = 0$ K, impurity atoms are not ionized, all of the valence electrons contribute to covalent bonds, the valence band is filled and there are no electrons in the conduction band. As the temperature rises, impurity atoms become ionized. At very low temperatures, $D \gg 1$, so the unity in the radicand in Eq. (9-62) may be neglected. Since $\sqrt{D} \gg 1$, too, the unity in the denominator may also be ignored. As a result, Eq. (9-62) reduces to

$$n_n = n_0 = \sqrt{(N_c N_d/2)}\exp[-(E_c - E_d)/2kT] \qquad (9\text{-}64)$$

Similarly, Eq. (9-63) may be re-written as

$$p_p = p_0 = \sqrt{(N_v N_a/2)}\exp[-E_a - E_v)/2kT] \qquad (9\text{-}65)$$

As is seen, the concentration of electrons in an N-type semiconductor due to ionization of donor atoms increases with rising temperature exponentially. In Fig. 9-17, this is shown by the straight line AB

whose slope (tan β) is proportional to the ionization energy of donors, $\Delta E_d = E_c - E_d$. In this temperature range, generation of electron-hole pairs is negligible, so, if there are no acceptor impurities, minority carriers, that is, holes, are almost absent in the semiconductor.

As the temperature approaches T_s known as the *saturation temperature*, practically all of the impurity atoms are ionized. Here, $D \ll 1$ and, as follows from Eq. (9-62), $n_n =$ (approx.) N_d. The region BC in Fig. 9-17 within which the electron concentration in an N-type semiconductor changes insignificantly is called the *saturation impurity-depletion range*.

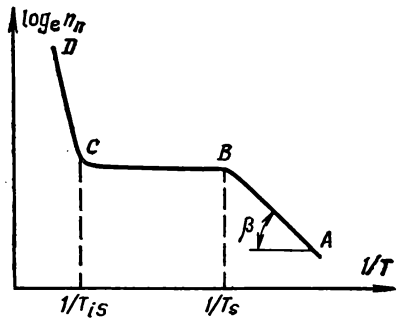


Fig. 9-17. Electron concentration in an N-type semiconductor as a function of temperature

The slight change in the concentration within the region BC is due to the fact that generation of electron-hole pairs is still insignificant. As the temperature rises further, generation becomes more intensive, and the concentration in the region CD increases according to Eq. (9-56). As the temperature changes from T_s to T_i, minority carriers (holes) are produced concurrently with majority carriers. For extrinsic semiconductors, Eq. (9-55) has the following form

$$n_n p_n = p_p n_p = n_i^2 \tag{9-66}$$

Hence, using Eq. (9-56), we get

$$p_n = n_i^2/N_d = (N_c N_v/N_d) \exp(-\Delta E_f/kT) \tag{9-67}$$

Thus, when the temperature of a semiconductor varies within the operating range, $T_i > T > T_s$, the concentration of minority carriers increases, and this seriously affects some important parameters of semiconductor devices.

The Fermi Level in Extrinsic Semiconductors. Expressions for the Fermi level in an N-type semiconductor, E_{Fn}, or in a P-type semiconductor E_{Fp}, may be obtained by solving simultaneously Eqs. (9-62) and (9-50) or Eqs. (9-63) and (9-52). The formulas turn out to be rather cumbersome, but they can easily be simplified when required. For instance for low temperatures ($T < T_s$) at which partial ionization of impurity atoms takes place, the equations can be solved to give [22]

$$E_{Fn} = (E_c + E_d)/2 - (kT/2) \log_e (2N_c/N_d) \tag{9-68}$$

$$E_{Fp} = (E_v + E_a)/2 + (kT/2) \log_e (2N_v/N_a) \tag{9-69}$$

From this, it follows that at $T = 0$ K, the Fermi level in a N-type semiconductor is located midway between the bottom of the conduction band and the donor level

$$E_{Fn} = (E_c + E_d)/2 \tag{9-70}$$

and in a P-type semiconductor, it lies midway between the top of the valence band and the acceptor level

$$E_{Fp} = (E_v + E_a)/2 \tag{9-71}$$

As the temperature rises, E_{Fn} gradually descends from the bottom of the conduction band to the centre of the forbidden band, and E_{Fp}

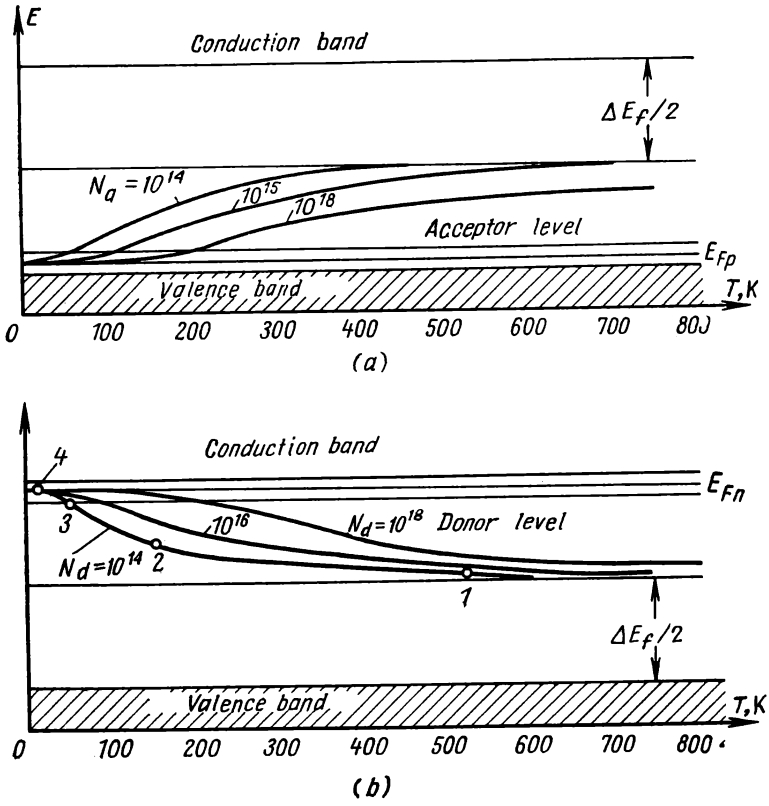


Fig. 9-18. Location of the Fermi level on the energy diagram of an extrinsic semiconductor as a function of temperature
(*a*) *P*-type semiconductor; (*b*) *N*-type semiconductor

moves from the top of the valence band. When the temperature reaches the saturation value, T_s, E_{Fn} becomes approximately equal to E_d, and E_{Fp} to E_a. Under the circumstances, about one third of the impurity atoms are ionized. The impurity atoms become ionized completely when E_{Fn} lies several $kT's$ below E_d, and E_{Fp} several $kT's$

above E_a. Most often, complete ionization occurs at a temperature of several tens of kelvins.

When the temperature varies from T_i to T_s, the Fermi level changes according to the following equations

$$E_{Fn} = E_c - kT \log_e (N_c/N_d) \tag{9-72}$$

$$E_{Fp} = E_v + kT \log_e (N_v/N_a) \tag{9-73}$$

As the temperature rises further, the Fermi level in extrinsic semiconductors shifts still more towards the centre of the forbidden band. Different positions of the Fermi level in extrinsic semiconductors with different impurity concentrations are illustrated in Fig. 9-18. Rising N_d and N_a result in higher T_s and T_i.

9-6. MOTION OF CHARGE CARRIERS. ELECTRICAL CONDUCTION IN SEMICONDUCTORS

Random Motion of Free Charge Carriers. In a semiconductor crystal, free particles move at random, and the probability that they will move in any direction is the same. The energy distribution of particles is described in terms of the Fermi-Dirac quantum statistics. In the absence of an external field, the densities of any two opposing electron streams are equal, that is, the current in the crystal is zero. The random motion of free particles is characterized by average statistical quantities, namely the mean free path, l_m, mean free time, t_m, and mean thermal velocity, v_m. The mean free path is the average distance that a particle travels between two collisions with other particles, collisions with vertices of the crystal lattice, and so on. Along the mean free path, the direction and value of particle velocity may be considered constant. The collisions may considerably change the velocity of a particle and its direction. The collisions of particles involve scattering, and the imperfections in the crystal lattice where particles experience scattering are called *scattering centres.*

In semiconductors, scattering centres may be neutral or ionized impurity atoms, vacancies, dislocations and other lattice defects. The physical nature of scattering differs from centre to centre. According to the nature of these processes, the mean free path for particles in semiconductors varies from 10^{-6} to 10^{-2} cm.

Drift. If at $T > 0$ K a semiconductor is subjected to an electric field with strength $\mathscr{E}$, the random motion of particles will be accompanied by an oriented motion. The ensemble of electrons each of which is moving at random will travel in the direction opposite to the field strength vector. This oriented motion, known as the *electron drift*, is, in fact, the transfer of charges under the action of an external field and underlies the concept of electrical conduction.

The drift velocity of electrons, v_d, may be defined as the mean velocity of particle transfer under the action of the electric field $\mathscr{E}$ over the time t_m

$$v_d = (e\mathscr{E}/m_n^*)\, t_m \qquad (9\text{-}74)$$

where $e\mathscr{E}/m_n^*$ is the acceleration gained by the electron in the field $\mathscr{E}$.

In relatively weak fields, v_d is small as compared with the mean velocity of the random motion. In an electric field 10 V/cm strong, v_d is approximately equal to 10^4 or 10^5 cm/s, and the thermal velocity at T = (approx.) 300 K is about 10^7 cm/s.

Drift Mobility. The directed motion is characterized in terms of the *drift mobility of the charge carriers*, μ, measured in $cm^2/V \cdot s$. It is defined as the average drift velocity per unit electric field, $\mathscr{E}$ = 1 V/cm:

$$\mu = v_d/\mathscr{E} \qquad (9\text{-}75)$$

Substituting Eq. (9-74) into the above expression gives for electrons

$$\mu_n = et_m/m_n^* \qquad (9\text{-}76)$$

Similarly, for holes

$$\mu_p = et_m/m_p^* \qquad (9\text{-}77)$$

The last expressions suggest that the mobility is inversely proportional to the effective mass of a particle. Since in the general case $m_n^* \neq m_p^*$, so $\mu_n \neq \mu_p$, (see Table 1-1). For example, for intrinsic germanium, $\mu_n/\mu_p = 2.1$, and for intrinsic silicon, $\mu_n/\mu_p = 2.8$.

Mobility as a Function of Temperature. The mobility is one of the most important parameters defining the motion of particles and, as a consequence, electrical conduction in semiconductors. In many respects, the mobility depends on the predominant scattering process. A plot of the relation $\log_e \mu = \varphi(1/T)$ typical of semiconductors is given in Fig. 9-19. At very low temperatures, thermal oscillations of the crystal lattice are small and most impurity atoms are not ionized. So, scattering mainly occurs at lattice defects and neutral impurity atoms. In this temperature interval, the mobility of particles increases with increasing temperature and markedly depends on the concentration of impurities, decreasing with increasing N. A further rise in temperature is accompanied by ionization of impurity atoms, so that scattering takes place at impurity ions and thermally oscillating lattice atoms. In the operating temperature range, the latter process becomes predominant, and the mobility

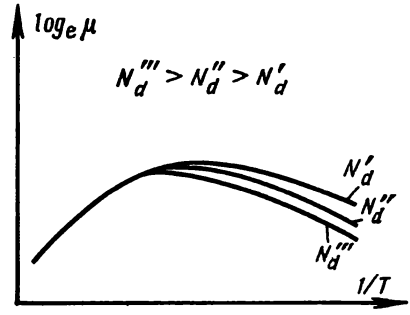


Fig. 9-19. Mobility as a function of temperature

decreases in proportion to T^{-c}, where c ranges between 1.6 and 2.6 for different semiconductors.

Electrical Conduction in Semiconductors. In a semiconductor, the drift of electrons and holes caused by the electric field $\mathscr{E}$ results in drift currents whose density for electrons is

$$j_{dn} = env_d = e\mathscr{E}n\mu_n \tag{9-78}$$

and for holes

$$j_{dp} = epv_d = e\mathscr{E}p\mu_p \tag{9-79}$$

Since electrons and holes carry unlike charges and drift in opposite directions under the action of the electric field, the resultant drift current density may be defined as

$$j_d = j_{dn} + j_{dp} = e\mathscr{E}(n\mu_n + p\mu_p) \tag{9-80}$$

The coefficient of proportionality between the current density and field strength is called the *conductivity*

$$\sigma = 1/\rho = e(n\mu_n + p\mu_p) \tag{9-81}$$

where ρ is the resistivity.

In an intrinsic semiconductor, $n_i = p_i$, so the drift current density is

$$j_d = j_{dn} + j_{dp} = e\mathscr{E}n_i(\mu_n + \mu_p) \tag{9-82}$$

As a rule, $\mu_n > \mu_p$, so $j_{dn} > j_{dp}$.

In the case of extrinsic semiconductors, the above expression is as follows: for N-type semiconductors

$$j_{d(n)} = e\mathscr{E}(n_n\mu_n + p_n\mu_p) \tag{9-83}$$

and for P-type semiconductors

$$j_{d(p)} = e\mathscr{E}(p_p\mu_p + n_p\mu_n) \tag{9-84}$$

In an N-type semiconductor, $n_n \gg p_n$, so $\mu_n > \mu_p$. Thus

$$j_{d(n)} = (\text{approx.})\, j_{dn} = (\text{approx.})\, e\mathscr{E}n_n\mu_n \tag{9-85}$$

For a P-type semiconductor, $p_p \gg n_p$ and, although $\mu_n > \mu_p$, we may take it that

$$j_{d(p)} = (\text{approx.})\, j_{dp} = (\text{approx.})\, e\mathscr{E}p_p\mu_p \tag{9-86}$$

In accordance with the above equations, we may define the conductivity as follows:

—for intrinsic semiconductors

$$\sigma_i = en_i(\mu_n + \mu_p) \tag{9-87}$$

—for N-type semiconductors

$$\sigma_n = (\text{approx.})\, en_n\mu_n \tag{9-88}$$

—for *P*-type semiconductors

$$\sigma_p = \text{(approx.)}\ ep_p\mu_p \qquad (9\text{-}89)$$

Electrical conduction due to generation of electron-hole pairs is called *intrinsic conduction*, and that due to ionization of impurity atoms is known as *extrinsic conduction*. On top of that, semiconductors

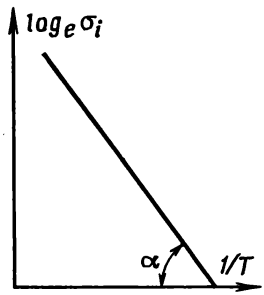


Fig. 9-20. Conductivity of an intrinsic semiconductor as a function of temperature

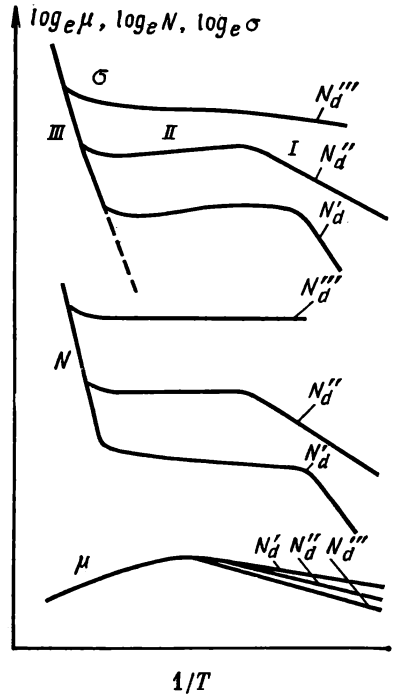


Fig. 9-21. Concentration and mobility of free particles and the conductivity of an *N*-type semiconductor versus temperature

possess *hole* (positive, or *P*-type) and *electron* (negative, or *N*-type) *conduction* mainly constituted by holes and electrons, respectively.

Recalling Eq. (9-56), we may show that

$$\log_e \sigma_i = f(-\Delta E_f/2kT) \qquad (9\text{-}90)$$

In the $\log_e \sigma_i - 1/T$ coordinates, this relation yields a straight line whose slope is proportional to $\Delta E_f/2$ (Fig. 9-20).

In an extrinsic semiconductor, the relationship $\sigma(T)$ has a more complex nature. To plot it (Fig. 9-21), we shall use the curves $\log_e n_n = f(1/T)$ and $\log_e \mu = \varphi(1/T)$ already constructed in Fig. 9-17 and 9-19, respectively.

Let us discuss the N_d'' curve. At low temperatures, the curve shows a well-defined region, *I*, representing extrinsic conduction. When the impurities are depleted (region *II*), the conduction may somewhat decrease due to a fall in μ. Portion *III* represents intrinsic conduction. When the concentration of impurities is sufficiently high

($N = N_d'''$), the curve has no portion *II*. This form of the $\log_e \sigma = f(1/T)$ relation is typical of highly-alloyed semiconductors, that is, semimetals, in which the impurity band and the nearest energy band overlap.

Motion of Particles under the Action of Strong Electric Fields. In electric fields whose strength does not exceed 5×10^2 V/cm, the mobility of particles does not depend on $\mathscr{E}$. As a consequence, the drift velocity of particles is proportional to the field strength, and the semiconductor obeys Ohm's law. The average drift velocity in weak electric fields is much smaller than the mean thermal velocity ($v_d \ll v_m$).

When the field strength exceeds 5×10^2 V/cm, the electron-hole gas behaves as if it were heated, and the particles become "hot" because their average energy corresponds to a temperature exceeding the temperature of the crystal lattice. Under the circumstances, the drift velocity approaches the mean thermal velocity ($v_d \gg 0.1\, v_m$) and becomes proportional to $\sqrt{\mathscr{E}}$, whereas the mobility decreases inversely proportional to $\sqrt{\mathscr{E}}$. Thus, the semiconductor has the properties of nonlinear resistance.

As the field strength rises further ($\mathscr{E} > 10^4$ V/cm), the average drift velocity becomes almost equal to the mean thermal velocity ($v_d \approx v_m \approx 10^7$ cm/s). The hot electrons exchange energy with the crystal lattice and give rise to optical phonons, that is, quanta of high-frequency oscillatory energy. Each time such a particle is produced, a considerable amount of energy is given by electron gas to the crystal lattice. Now, the average drift velocity is independent of the field strength. The current in the semiconductor crystal reaches saturation which is unaffected by variations in $\mathscr{E}$. The electric field strength at which this effect occurs is called *critical*. For most semiconductors, $\mathscr{E}_{cr}$ = (approx.) 10^4 V/cm.

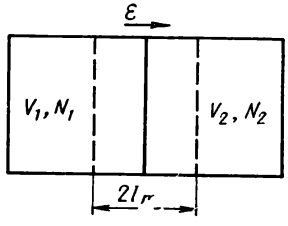


Fig. 9-22. Explaining diffusion

Diffusion of Particles. As is known from a course on physics, if we have two elementary cubic volumes filled with gas particles, one with the concentration N_1 and the other, N_2 (Fig. 9-22), the gas particles will diffuse through the interface between the two volumes. In both volumes, gas particles move at random with a mean thermal velocity of

$$v_m = \sqrt{3kT/m} \tag{9-91}$$

and the probability that they will move in any direction is the same. So approximately one sixth of all particles in each volume will travel at velocities whose vectors are directed toward the inter-

face. Since $N_1 > N_2$, the interface will pass more particles from V_1 to V_2 than in the reverse direction. This directed motion of particles due to the difference in concentration between the two elementary volumes is called *diffusion*.

Thus, the diffusion current is

$$J = 1/6\ (N_1 - N_2)\ v_m \tag{9-92}$$

The particles diffusing into the adjacent volume interact with its particles, on the average, within the mean free path, l_m. So, owing to diffusion, the concentrations of particles differ from equilibrium concentrations within l_m of the interface. Consequently, we may write for the concentration gradient

$$dN/dx = (N_1 - N_2)\ 2l_m \tag{9-93}$$

Substituting this into Eq. (9-92) gives

$$J = (1/3)\ v_m l_m\ dN/dx \tag{9-94}$$

The coefficient of the concentration gradient in this expression is called the *diffusion constant* (or *coefficient*)

$$D = (1/3)\ v_m l_m \tag{9-95}$$

Now, let us assume that the volume V_1 contains negative mobile particles, electrons, whose space charge is neutralized by the positive charge of immobile particles. As a whole, the volume V_1 is electrically neutral. The condition for charge neutrality is also met in the volume V_2, but here the mobile particles, holes, are charged positively. When the partition is removed, the electrons will diffuse from V_1 to V_2, and the holes from V_2 to V_1. This motion is accompanied by the redistribution of electric charges, that is, an electric field $\mathscr{E}$ appears. The field will force the particles to drift in the direction opposite to their diffusion. Under conditions of equilibrium, the drift and diffusion current densities should be equal (the Fermi energy for the whole system is E_F = constant). In other words, the time it takes particles to diffuse for the mean free path should be equal to the time it takes them to drift for the same distance. In an electric field, particles are in a uniformly accelerated motion at the average drift velocity defined by Eq. (9-74)

$$v_d = (e\mathscr{E}/m)\ t_m$$

Hence, recalling that $v_d = \mu\mathscr{E}$, we may write for the mean free time

$$t_m = m\mu/e \tag{9-96}$$

The diffusing particles travel the same distance in

$$t = l_m/v_m \tag{9-97}$$

Because the opposing streams balance each other, these times are equal and, consequently,

$$l_m = m\mu v_m/e \tag{9-98}$$

Substituting this expression into Eq. (9-95) and using Eq. (9-91), we get

$$D = \mu kT/e \tag{9-99}$$

This equation is known as the *Einstein relationship between mobility and diffusion constant.*

Diffusion Current Density. On the basis of Eqs. (9-94), (9-95) and (9-99), we may define the diffusion current density

$$j_{Dn} = eD_n\, dn/dx \tag{9-100}$$

Here, the subscript D stands for the diffusion current, and

$$D_n = \mu_n kT/e \tag{9-101}$$

is the *diffusion constant of electrons.*

Similarly, for holes we have

$$j_{Dp} = -eD_p\, dp/dx \tag{9-102}$$

where

$$D_p = \mu_p kT/e \tag{9-103}$$

Since electrons and holes differ in mobility, their diffusion constants D_n and D_p are also different.

These constants uniquely defined by the mobility of particles as well as the mobility μ itself, depend on temperature and the concentration of particles. The character of these relations has already been discussed (see Sec. 9-6).

For a semiconductor containing both free electrons and holes, the resultant diffusion current density may be defined as

$$j_D = j_{Dn} + j_{Dp} = e(D_n\, dn/dx - D_p\, dp/dx) \tag{9-104}$$

Total Current Density. The current flowing through most semiconductor devices is the sum of the drift and diffusion currents due to free charge particles. So, recalling Eqs. (9-80) and (9-104), the total current density in the general case may be defined as

$$j = j_d + j_D = en\mu_n \mathscr{E} + ep\mu_p \mathscr{E} + eD_n\, dn/dx - eD_p\, dp/dx \tag{9-105}$$

All terms in this expression contain the concentrations of positive or negative particles which, in the general case, depend on the position and time, that is, $p = f(x, t)$ and $n = \varphi(x, t)$. As already noted, the drift and diffusion of particles are accompanied by generation of electron-hole pairs, ionization of impurities, and their recombination. The intensity of these processes may be different for various parts of a semiconductor crystal. In any case, however, all these processes in either the entire crystal or only a part thereof

should obey a fundamental physical law, the law of conservation of electric charge

$$-\partial\rho/\partial t = \operatorname{div} j \tag{9-106}$$

where ρ is the volume charge density.

In other words, a change in the volume charge density in any volume per unit time should be equal to the number of elementary electric charges leaving (or entering) that volume.

Continuity Equation. If a semiconductor contains free unlikely charged particles and also taking into consideration that generation of electron-hole pairs and their recombination take place in it, the law of conservation of charge given by Eq. (9-106) may be described by what is known as *continuity equations*

$$\partial p/\partial t = -(p - p_0)/\tau_p - (1/e)\operatorname{div} j_p + G_p \tag{9-107}$$

$$\partial n/\partial t = -(n - n_0)/\tau_n + (1/e)\operatorname{div} j_n + G_n \tag{9-108}$$

Here, the first terms on the right-hand sides describe recombination in accordance with Eq. (9-26) (p and n are the non-equilibrium concentrations, p_0 and n_0 are the equilibrium concentrations, and τ_p and τ_n are the lifetimes of the non-equilibrium charge carriers); G_p and G_n characterize the generation rates of electrons and holes due to external factors. The second terms on the right-hand sides may be re-written using Eq. (9-105) thus

$$(1/e)\operatorname{div} j_p = (1/e)\,\partial\,(j_{dp} + j_{Dp})/\partial x$$
$$= \mu_p \mathscr{E}\,\partial p/\partial x + p\mu_p\,\partial\mathscr{E}/\partial x - D_p\,\partial^2 p/\partial x^2 \tag{9-109}$$

$$(1/e)\operatorname{div} j_n = (1/e)\,\partial\,(j_{dn} + j_{Dn})/\partial x$$
$$= \mu_n \mathscr{E}\,\partial n/\partial x + n\mu_n\,\partial\mathscr{E}/\partial x + D_n\,\partial^2 n/\partial x^2 \tag{9-110}$$

Substituting these expressions into Eqs. (9-107) and (9-108) gives the continuity equations in their general form

$$\partial p/\partial t = -(p - p_0)/\tau_p + D_p\,\partial^2 p/\partial x^2 - \mu_p \mathscr{E}\,\partial p/\partial x - p\mu_p\,\partial\mathscr{E}/\partial x + G_p \tag{9-111}$$

$$\partial n/\partial t = -(n - n_0)/\tau_n + D_n\,\partial^2 n/\partial x^2 + \mu_n \mathscr{E}\,\partial n/\partial x + n\mu_n\,\partial\mathscr{E}/\partial x + G_n \tag{9-112}$$

As already noted, the first and the last terms in these equations describe the rates of generation and recombination of carriers, the second term their diffusion, and the third and the fourth terms their drift.

The general solution of these second-order differential equations in partial derivatives is rather difficult to find. In most practical cases, however, Eqs. (9-111) and (9-112) may be simplified.

One of such cases, where it is necessary to find the recombination time of a non-equilibrium concentration in the absence of drift and diffusion, has already been discussed in Sec. 9-3 [see Eqs. (9-26) and (9-27)].

Another case, rather common in the theory of semiconductor devices, involves the transformation of the continuity to diffusion equations. Suppose that in a volume ΔV of an N-type semiconductor some external factors result in a non-equilibrium concentration of holes, p, which is Δp over and above the equilibrium concentration, p_0. Also, let us assume that Δp is much smaller than the equilibrium concentration of electrons, n_0, so that the electrons which have entered ΔV to balance the positive charge can affect the redistribution of charges in the semiconductor only insignificantly, and the field $\mathscr{E}$ is negligible. Then the resulting concentration gradient will cause holes and electrons to diffuse from the volume ΔV and gradually recombine. Now, in the steady state and provided that $G_p = G_n = 0$, the continuity equations may be re-written as follows

$$D_p \partial^2 p/\partial x^2 - (p - p_0)/\tau_p = 0 \qquad (9\text{-}113)$$

$$D_n \partial^2 n/\partial x^2 - (n - n_0)/\tau_n = 0 \qquad (9\text{-}114)$$

Solving Eq. (9-113), we may obtain a law by which the concentration of minority carriers changes with x reckoned from the region where the non-equilibrium concentration is formed. If the semiconductor crystal is sufficiently large, the boundary conditions have the form: $p\ (x = 0) = p_0 + \Delta p$ and $p\ (x \to \infty) = p_0$.

The solution for Eq. (9-113) is

$$p\ (x) = p_0 + \Delta p \exp\ (-x/L_p) \qquad (9\text{-}115)$$

where the quantity

$$L_p = \sqrt{D_p \tau_p} \qquad (9\text{-}116)$$

called the *diffusion length*, is equal to the average distance travelled by non-equilibrium carriers during their lifetime.

The average velocity

$$v_D = L_p/\tau_p = D_p/L_p \qquad (9\text{-}117)$$

at which the diffusing carriers travel the path L_p is called the *diffusion velocity*.

A plot of the concentration of carriers as a function of x, $p\ (x)$, is shown in Fig. 9-23*a*.

Equation (9-113) has a different solution, if the length of the crystal, w, is small ($w \ll L_p$) and there are no nonrecombined carriers left in the $x = w$ plane:

$$p\ (x) = p_0 + \Delta p \{\sinh\ [(w - x)/L_p]/\sinh\ (w/L_p)\} \qquad (9\text{-}118)$$

Assuming that $w \ll L_p$, the trigonometric functions may be replaced by their arguments

$$p(x) = p_0 + \Delta p\,(1 - x/w) \tag{9-119}$$

In this case, the carrier distribution is linear (Fig. 9-23*b*).

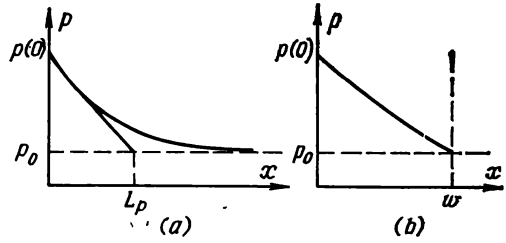


Fig. 9-23. Concentration of carriers as a function of x in the case of diffusion (*a*) at $w > L_p$; (*b*) at $w \ll L_p$

The diffusion velocity for this case may be derived using Eq. (9-102) for the diffusion current density and also keeping in mind that $j = \rho e v$. The simultaneous solution of these equations gives

$$v_D = (1/p)\,D dp/dx \tag{9-120}$$

Determining the concentration gradient dp/dx by differentiating Eq. (9-119) with respect to x, and substituting the result into Eq. (9-120), we get the average diffusion time for

$$t_D = w/v_D = w^2/2D \tag{9-121}$$

Chapter Ten

JUNCTION EFFECTS

10-1. SEMICONDUCTOR JUNCTIONS

Definition. A semiconductor junction is a region of transition between two semiconductor regions with different types of conduction or different conductivities.

Types of Junctions. A junction composed of two semiconductor regions with P-type and N-type conductions is called the *electron-hole junction*, or more often, the *P-N junction.*

If a junction is made up of semiconductor regions having the same type of conduction but differing in conductivity, we speak of *N-N* or *P-P junctions.*

One of the regions taking part in a junction may be a metal. In this case, we speak of a *metal-semiconductor junction.*

If a junction is composed of two dissimilar semiconductors (having different energy gaps), for example, germanium and gallium arsenide, it is called a *heterogeneous junction.*

Junctions cannot be fabricated by a simple mechanical contact between two semiconductor crystals, because their surfaces are contaminated with atoms of other materials, semiconductor oxides, and the like. There are several methods of junction manufacture, such as alloying of a part of an N-type semiconductor crystal with acceptor impurities by their diffusion from a gaseous or liquid medium containing atoms of the required impurity (junctions fabricated by this technique are known as *diffused junctions*). Junctions may also be manufactured by alloying a metal or an alloy containing acceptor or donor impurities in a semiconductor (*alloy junctions*), and so on.

10-2. THE SYMMETRICAL P-N JUNCTION

Physical Processes. Let a *P-N* junction be composed in a semiconductor crystal one part of which is alloyed with acceptor impurities (P-region) and the other with donor impurities (N-region) with the impurity concentrations in both regions being equal, $N_a = N_d$.

Let us also suppose that at room temperature ($T = 300$ K) practically all of the impurity atoms are ionized, and both regions contain a certain amount of minority carriers due to generation of electron-hole pairs. Thus, the P-region has the equilibrium concentrations of majority carriers, p_{p0}, and minority, carriers, n_{p0}. Similarly, the equilibrium concentrations of majority and minority carriers in the N-region are n_{n0} and p_{n0}, respectively. In the P-region, $n_{p0} \ll p_{p0}$, and in the N-region, $p_{n0} \ll n_{n0}$. Since $N_a = N_d$, the

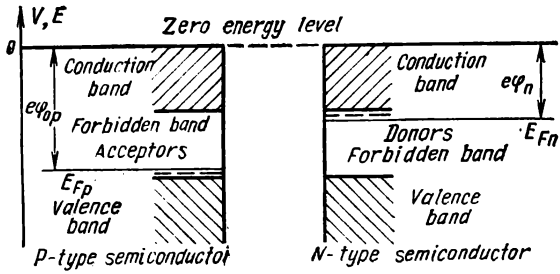


Fig. 10-1. Energy diagrams of semiconductors out of contact

concentrations of majority and minority carriers in both regions are equal: $p_{p0} = n_{n0}$ and $n_{p0} = p_{n0}$, and the condition for the charge neutrality specified by Eqs. (9-29) and (9-31) holds for both regions. This is a *symmetrical P-N junction.* Figure 10-1 shows separately the energy-band diagrams of both crystal regions. In a real junction, it is impossible to draw a distinct line between the P- and N-regions. As a first approximation, however, let the junction be ideally abrupt (Fig. 10-2*a*), and the regions be separated by a plane known as the *P-N boundary* which is taken as the origin of the x-coordinate. The concentrations of like carriers on both sides of this boundary are different: $p_{p0} \gg p_{n0}$ and $n_{n0} \gg n_{p0}$, so there is a nonzero concentration gradients of electrons and holes:

$$dp/dx \neq 0 \qquad (10-$$

$$dn/dx \neq 0 \qquad (10$$

The difference in concentration leads to the diffusion of partic holes diffuse from the P- to N-region, and electrons move in opposite direction. It should be emphasized that this motion nothing to do with the mutual repulsion of like, and attracti unlike, charges. The diffusion takes place only owing to the ference in particle concentration on both sides of the P-N bou

The diffusion is constituted only by the free particles o side and within the average diffusion length of the bound crossing the boundary, holes find themselves in the N-regi they gradually recombine with electrons whose concent

that region is high. Electrons entering the P-region recombine with holes in much the same way. This upsets the condition for charge neutrality on either side and within the distance l of the boundary. The P-region has uncompensated negative charges due to immobile acceptor ions, and the N-region, uncompensated positive charges due to immobile donor ions. This double electric charge layer develops an electric field (Fig. 10-2*d*) whose strength, $|\mathscr{E}_c|$, increases with increasing diffusion of majority carriers and, as a consequence, with increasing space charges due to immobile impurity ions. The field strength vector is directed so that it opposes the diffusion of majority carriers which is the very process that has given rise to the field itself. As the field $\mathscr{E}_c$ gains in strength, the diffusion of majority carriers across the boundary decreases, and an increasing number of holes and electrons are swept by this field away from the boundary and back to their original regions.

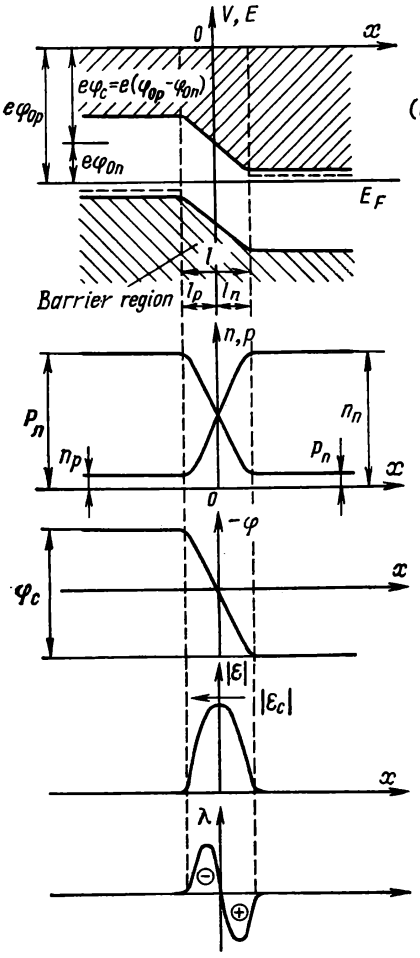


Fig. 10-2. Symmetrical P-N junction
(*a*) energy diagram of the junction; (*b*) concentration of mobile carriers; (*c*) potential distribution; (*d*) field strength; (*e*) concentration of immobile carriers

At the same time, the field $\mathscr{E}_c$ is responsible for the drift of minority carriers. Holes move from the N- to P-*region*, and electrons drift in the opposite direction. As a result, the majority carriers—holes and electrons—move in the opposite directions, but since they carry unlike charges, they contribute the positive and negative components to the same diffusion current flowing in the same direction with holes

$$j_D = j_{Dp} + j_{Dn} \tag{10-3}$$

Similarly, the minority-carrier drift current density may be defined as

$$j_d = j_{dp} + j_{dn} \tag{10-4}$$

The *P-N* junction will attain a state of equilibrium when the field intensity is such that the diffusion current is completely balanced by the drift current and the total current across the junction is zero

$$j = j_D + j_d = 0 \tag{10-5}$$

Using Eq. (9-105), we may write the above expression in the general form which holds for any cross section of a semiconductor crystal

$$j = j_D + j_d = e\,(D_n\,dn/dx - D_p\,dp/dx) + e\mathscr{E}_c\,(n\mu_n + p\mu_p) = 0 \tag{10-6}$$

or separately for the hole and electron components

$$j_n = j_{Dn} + j_{dn} = e\,(D_n\,dn/dx + \mathscr{E}_c n\mu_n) = 0 \tag{10-7}$$

$$j_p = j_{Dp} + j_{dp} = e\,(-D_p\,dp/dx + \mathscr{E}_c p\mu_p) = 0 \tag{10-8}$$

As the total current is zero, the entire system should have a common Fermi level (Fig. 10-2*a*).

In contrast to a system containing electrically neutral particles, in our case the concentrations of particles on either side of the boundary do not become equal under conditions of equilibrium. At equilibrium, the concentration gradients of like carriers are nonzero ($dn/dx \neq 0$ and $dp/dx \neq 0$), but the diffusion of particles is opposed by a potential barrier which appears near the boundary and whose value is equal to the contact potential difference (Fig. 10-2*c*)

$$\varphi_c = \varphi_{0p} - \varphi_{0n} \tag{10-9}$$

Depletion (Barrier) Layer. Figure 10-2*a* through *e* shows the energy-band diagram of an abrupt symmetrical *P-N* junction at equilibrium and, approximately, the manner in which the mobile-carrier concentrations, n and p, potential, φ, field strength, $\mathscr{E}_c$, and concentration of immobile charges, λ, vary with x. The region $l = l_p + l_n$ shown by the dashed line within which the above quantities vary is called the *depletion*, or *barrier layer.*

Beyond the depletion layer, the energy-band diagrams of the *P*- and *N*-regions are not distorted. Here, the semiconductor is electrically neutral, and no electric field exists.

As already noted, mobile charge carriers leaving both regions upset electrical neutrality and give rise to uncompensated negative and positive space charges of immobile ionized acceptor and donor atoms.

The barrier layer is depleted of mobile charge carriers, so its resistance must be considerably higher than the resistance of the regions outside it. Actually, the barrier layer retains a certain number of mobile carriers, because electrons and holes having some thermal energy penetrate into the depletion layer and are reflected by the field $\mathscr{E}_c$. Also, generation of electron-hole pairs and their recombination may occur in this layer. These phenomena will be discussed

later in Chapter 11 which deals with the physical processes in real semiconductor devices. For the time being we shall take it that space charges in the depletion layer are formed only by ionized impurity atoms. Using such an idealized picture of a *P-N* junction, we shall define the main physical quantities.

The Height of the Potential Barrier. At a *P-N* junction, the height of the potential barrier is equal to the contact potential difference, $\varphi_c = \varphi_{0p} - \varphi_{0n}$. The work function of semiconductors is reckoned from the Fermi level, so $e\varphi_c = E_{Fp} - E_{Fn}$. Using Eqs. (9-72) and (9-73) which define the Fermi level for *P*- and *N*-type semiconductors, we may write

$$e\varphi_c = \Delta E_f - kT \log_e (N_v N_c / N_a N_d) \tag{10-10}$$

Substituting from Eq. (9-67)

$$N_v N_c = n_i^2 \exp (\Delta E_f / kT) \tag{10-11}$$

gives

$$e\varphi_c = kT \log_e (N_a N_d / n_i^2) \tag{10-12}$$

Putting $N_a = p_{p0}$ and $N_d = n_{n0}$, we get

$$e\varphi_c = kT \log_e (n_{n0} p_{p0} / n_i^2) \tag{10-13}$$

or, recalling that $n_i^2 = n_{n0} p_{n0} = p_{p0} n_{p0}$,

$$e\varphi_c = kT \log_e (p_{p0} / p_{n0}) = kT \log_e (n_{n0} / n_{p0}) \tag{10-14}$$

These expressions make the height of the potential barrier at the *P-N* junction very easy to determine. For example, for germanium, $n_i = p_i = 10^{13}$ cm^{-3}. If the impurity concentration in germanium is $N_a = N_d = 10^{16}$ cm^{-3} and, as a consequence, $p_{p0} = n_{n0} = 10^{16}$ cm^{-3}, then, according to Eq. (9-66), $n_{p0} = p_{n0} = 10^{10}$ cm^{-3}. Hence, the height of the potential barrier at $T = 300$ K according to Eq. (10-14) is about 0.36 eV.

Relation between the Concentrations on Either Side of the P-N Junction. This we may easily deduce by taking an antilogarithm of Eq. (10-14)

$$p_{p0} = p_{n0} \exp (e\varphi_c / kT) \tag{10-15}$$

and

$$n_{n0} = n_{p0} \exp (e\varphi_c / kT) \tag{10-16}$$

The Electric Field. The field $\mathscr{E}_c$ may be defined by means of Poisson's equation

$$\partial^2 \varphi / \partial x^2 = -\lambda (x) / \varepsilon \varepsilon_0 \tag{10-17}$$

As defined above, the space charges at the *P-N* junction are formed by ionized impurity atoms, so, for the depletion layer in the

P- and N-regions we may re-write Eq. (10-17) as follows

$$\partial^2\varphi/\partial x^2 = -eN_a/\varepsilon\varepsilon_0 \tag{10-18}$$

$$\partial^2\varphi/\partial x^2 = eN_d/\varepsilon\varepsilon_0 \tag{10-19}$$

Integrating these equations going from l_p to 0 and from 0 to l_n, respectively, gives

$$\mathscr{E}_c(x) = -eN_a(x+l_p)/\varepsilon\varepsilon_0 \tag{10-20}$$

$$\mathscr{E}_c(x) = eN_d(x-l_n)/\varepsilon\varepsilon_0 \tag{10-21}$$

The field strength $|\mathscr{E}_{c\text{-max}}|$ can easily be determined by putting $x = 0$ in any of the above expressions

$$|\mathscr{E}_{c\text{-max}}| = eN_a l_p/\varepsilon\varepsilon_0 = eN_d l_n/\varepsilon\varepsilon_0 \tag{10-22}$$

The Width of the Depletion Layer. Integrating Eqs. (10-18) and (10-19) twice between the same limits and adding together the results, we get the following expression for the height of the potential barrier

$$\varphi_c = \mathscr{E}_{c\text{-max}}(l_p + l_n)/2 \tag{10-23}$$

Putting $l_p + l_n = l$ and recalling Eq. (10-22), we may write for the width of the depletion layer

$$l = \sqrt{(2\varepsilon\varepsilon_0\varphi_c/e)(N_a+N_d)/N_aN_d} \tag{10-24}$$

It is an easy matter to find the width of the depletion layer, provided the impurity concentrations are known. For example, for silicon, $\varepsilon = 12$ (see Table 9-1). Putting $\varphi_c = 1$ V and $N_a = N_d = 10^{15}$ cm^{-3}, we get $l =$ (approx.) 1 μm. As the concentration of impurities decreases, the depletion layer becomes narrower: at $N_a = N_d = 10^{17}$ cm^{-3}, the width of the depletion layer is $l =$ = (approx.) 0.1 μm. For a P-N junction in a germanium crystal. the depletion layer is wider, because for germanium $\varepsilon = 16$.

10-3. FORWARD AND REVERSE BIASING

Forward Biasing. Consider the application of a potential difference to a P-N junction as shown in Fig. 10-3*a*. The P side has positive polarity, and the N side negative. This is called *forward biasing*.

As before, we shall assume that the resistance of the barrier region is much higher than those of the P- and N-regions, so all of the external voltage is applied directly to the transition region. The external voltage brings the potential barrier at the junction down to $\varphi_c - V$ (Fig. 10-3*b*), the equilibrium is upset ($j_D > j_d$), and a diffusion current due to holes moving from the P- to N-region and electrons moving in the opposite direction flows through the junction.

In other words, forward biasing causes the diffusion of particles across the junction to the region where they constitute a minority (holes are driven to the N-region, and electrons to the P-region); this is called the *injection of minority carriers.*

As the potential barrier goes down, the field strength $\mathscr{E}_c$ in the depletion layer decreases, and so does the width of the barrier re-

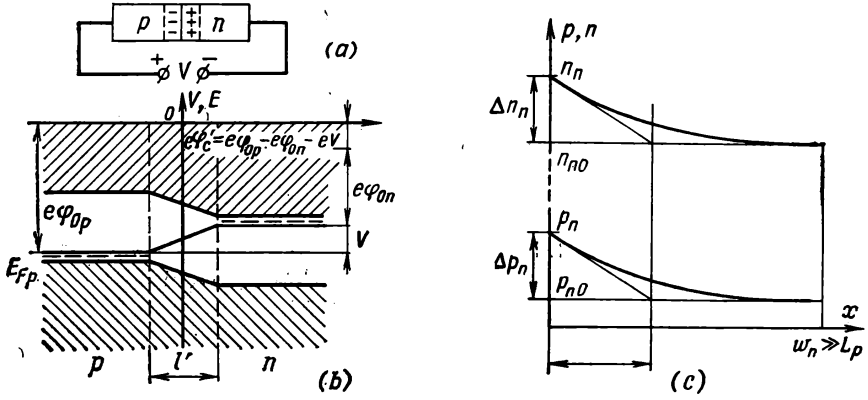


Fig. 10-3. (*a*) Forward-biased P-N junction; (*b*) its energy diagram; (*c*) concentrations of charge carriers

gion. Now Eq. (10-24) should contain the new value of the height of the potential barrier

$$l' = \sqrt{[2\varepsilon\varepsilon_0 (\varphi_c - V)/e] (N_a + N_d)/N_a N_d} \qquad (10\text{-}25)$$

The injection gives rise to non-equilibrium concentrations of minority carriers near the boundaries of the barrier region: holes in the N-region ($p_n > p_{n0}$) and electrons in the P-region ($n_p > n_{p0}$). The lower plot in Fig. 10-3*c* represents the concentration of holes in the N-region as a function of x (x is reckoned from the boundary of the barrier region). When $w_n < L_p$, the hole concentration gradient is

$$dp/dx = (p_n - p_{n0})/w_n = \Delta p_n/w_n \qquad (10\text{-}26)$$

and the electron concentration gradient in the P-region is

$$dn/dx = (n_p - n_{p0})/w_p = \Delta n_p/w_p \qquad (10\text{-}27)$$

Here, w_n and w_p are the lengths of the N- and P-regions, respectively, as reckoned from the boundaries of the barrier region to the terminals.

As a result, holes diffuse from the junction deep into the N-region where they gradually recombine with electrons, and electrons diffuse into the P-region where they gradually recombine with holes. At the outer boundaries of the semiconductor regions, the concentrations of the minority carriers attain a state of equilibrium, n_{p0} and p_{p0}.

According to Eqs. (9-100) and (9-102), the diffusion current densities due to the minority carriers injected into the P- and N-regions across the junction depend on the concentration gradients which, in this case, are defined by Eqs. (10-26) and (10-27). Let us discuss the main quantities defining the injection of minority carriers and their subsequent behaviour.

Relation between the Equilibrium and Non-Equilibrium Concentrations at the P-N Junction Boundaries. The concentrations p_n and n_p may be defined by Eqs. (10-15) and (10-16), if we substitute the height of the reduced potential barrier, $\varphi_c - V$, for φ_c

$$p_{p0} = p_n \exp(e\varphi_c/kT) \exp(-eV/kT) \tag{10-28}$$

$$n_{n0} = n_p \exp(e\varphi_c/kT) \exp(-eV/kT) \tag{10-29}$$

Substituting p_{p0} and n_{n0} from Eqs. (10-15) and (10-16), gives

$$p_n = p_{n0} \exp(eV/kT) \tag{10-30}$$

$$n_p = n_{p0} \exp(eV/kT) \tag{10-31}$$

From this we infer that the concentrations of minority carriers at the junction boundaries exponentially increase with increasing external voltage V.

Injection Level. This quantity describes relative variations in the concentration of injected minority carriers at the junction boundaries

$$\delta = \Delta p_n/n_{n0} = \Delta n_p/p_{p0} \tag{10-32}$$

As is seen from this expression, the injection level is the ratio of the increment in the concentration of minority carriers caused by injection to the equilibrium concentration of majority carriers. The fact that Δp_n and Δn_p are compared with the concentrations of majority carriers can be explained as follows. At the junction boundaries, the injected minority carriers form space charges. To balance them, majority carriers flow from both regions to the junction. This gives rise to excess concentrations of both minority and majority carriers at the junction. In the N-region, Δn_n = (approx.) Δp_n, and in the P-region, Δp_p = (approx.) Δn_p (Fig. 10-3*c*).

If $\Delta n_n \ll n_n$ and $\Delta p_p \ll p_p$, the redistribution of majority carriers between the regions does not practically disturb charge neutrality (the electric field developed in the regions is weak). In this case, the regions are called *quasi-neutral*, and the injection level is said to be *low* ($\delta \ll 1$).

If the increments in the concentrations of injected carriers are comparable with the concentrations of majority carriers, that is, Δp_n = (approx.) Δn_n = (approx.) n_n and Δn_p = (approx.) Δp_p = = (approx.) p_p, then the injection level is considered to be *medium* ($\delta \approx 1$); if $\delta > 1$; it is said to be *high*. In these two cases, the redi-

stribution of majority carriers between the semiconductor regions gives rise to a considerable internal electric field, because the majority carriers moving away from the depth of the regions leave behind uncompensated charges constituted by immobile impurity ions. So, the electric charge neutrality of the regions is disturbed. These cases will be discussed later with reference to real semiconductor devices. Meanwhile, when discussing the physical processes in the *P-N* junction, we shall deem the injection level to be low ($\delta \ll 1$), and the semiconductor regions beyond the junction boundaries electrically neutral.

Variations in the Concentrations of Minority Carriers. In the semiconductor regions beyond the junction boundaries, these may be described with the help of Eqs. (9-115) and (9-119). So long as the *P*- and *N*-regions exceed the diffusion lengths, $w_n > L_p$ and $w_p > L_n$, the concentrations of minority carriers vary exponentially (see Figs. 9-23*a* and 10-3*c*)

$$p(x) = p_{n0} + \Delta p \exp(-x/L_p) \tag{10-33}$$

$$n(-x) = n_{p0} + \Delta n \exp(x/L_n) \tag{10-34}$$

If $w_n \ll L_p$ and $w_p \ll L_n$, the concentrations change linearly

$$p(x) = p_{n0} + \Delta p(1 - x/w_n) \tag{10-35}$$

$$n(-x) = n_{p0} + \Delta n(1 + x/w_p) \tag{10-36}$$

As is seen, in either case, the concentrations of minority carriers at the outer boundaries of a semiconductor crystal are equal to the equilibrium concentrations. In the former case, that is, when $x \to \pm\infty$, $p(x) = p_{n0}$ and $n(-x) = n_{p0}$; the same holds for the latter case provided $x = w_n$ and $-x = w_p$.

If we know the manner in which the concentrations of minority carriers vary, we may determine the diffusion current density due to the carriers injected across the *P-N* junction. But first let us examine the physical processes in the *P-N* junction and adjacent regions when the positive terminal of the external voltage source is connected to the *N* side and the negative terminal to the *P* side, which is called *reverse biasing*.

Reverse Biasing. When a potential difference is applied to the *P-N* junction as shown in Fig. 10-4*a*, the junction is said to be *reverse-biased*. Under the circumstances, the potential barrier rises to $\varphi_c + |V|$ (Fig. 10-4*b*) and the junction is no longer at equilibrium. In contrast to forward biasing, however, the diffusion current density across the *P-N* boundary due to majority carriers is decreased ($j_D < j_d$), and a current due to minority carriers will flow through the junction, that is, holes will flow from the *N*- to *P*-regions, and electrons in the opposite direction.

In this case, the width of the barrier region increases

$$l'' = \sqrt{[2\varepsilon\varepsilon_0(\varphi_c + |V|)/e](N_a + N_d)/N_a N_d} \qquad (10\text{-}37)$$

and so does the electric field strength $|\mathscr{E}_c|$ at the P-N junction. The increased field $\mathscr{E}_c$ causes minority carriers to leave the semiconductor regions adjacent to the junction, so their concentrations at the junction goes down nearly to zero (Fig. 10-4*c*). This produces

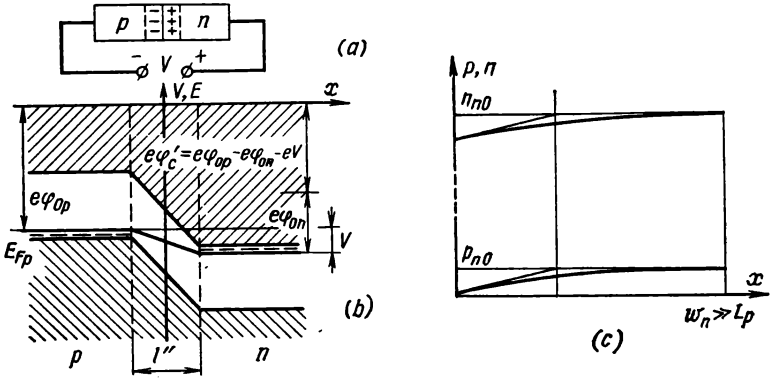


Fig. 10-4. (*a*) Reverse-biased P-N junction; (*b*) its energy diagram; (*c*) concentrations of charge carriers

minority-carrier concentration gradients near the P-N junction, and the carriers begin to diffuse from the semiconductor regions to the junction. This is called the *extraction of minority carriers.*

As the reverse voltage rises, the depletion layer becomes wider according to Eq. (10-37), the field $\mathscr{E}_c$ increases, but the current across the junction remains practically the same. The point is that the minority-carrier concentration gradients near the P-N junction do not change with increasing V, because n_{p0}, p_{n0}, L_p and L_n are independent of this voltage. For this reason, the current flowing across the reverse-biased P-N junction is called the *reverse saturation current*, I_0. This current may increase only with increasing concentrations, n_{p0} and p_{n0}, that is, with rising temperature. That is why it is also called the *thermal current.* When discussing real semiconductor devices in Chapter 11, we shall see that I_0 depends not only on the intensity of thermal carrier generation. For the time being we shall neglect the effect that generation and recombination of carriers, leakage current and some other factors have on the reverse current.

The Volt-Ampere Characteristic of an Idealized P-N Junction. In order to relate the current I flowing across a P-N junction to the external voltage V (the volt-ampere characteristic of the junction), let us use the continuity equations (9-111) and (9-112) which hold

for any cross section of a semiconductor crystal containing a P-N junction. We shall limit ourselves to solving Eq. (9-111) for holes which are minority charge carriers in the N-region. The final results may easily be extended to cover the P-region.

Under conditions of a low injection level, the electric field in the N-region is very weak. So we may deem the terms of Eq. (9-111) containing $\mathscr{E}$ equal to zero. Generation of charge carriers may also be neglected ($G_p = 0$). Thus, in the steady state, Eq. (9-111) may be reduced to an equation similar to Eq. (9-113)

$$D_p \partial^2 p/\partial x^2 - \Delta p_n/\tau_p = 0 \qquad (10\text{-}38)$$

or, taking into consideration that $L_p^2 = D_p \tau_p$, Eq. (10-38) may be re-written as

$$\partial^2 p/\partial x^2 - \Delta p_n/L_p^2 = 0 \qquad (10\text{-}39)$$

Let us specify the conditions under which this equation should be solved. We shall assume that the width of the barrier region is rather small ($l \to 0$) so that the boundary between the barrier region and the N-region is at $x = 0$. Also, we shall neglect the likely generation and recombination of carriers at the junction. We shall also assume that the voltage V varies so that the injection level is always low ($\delta \ll 1$) and, as a consequence, the N-region is electrically neutral. As before, the resistance of the depletion layer will be taken to be much higher than the volume resistances of the P- and N-regions and, consequently, almost all of the external voltage V will drop at the P-N junction.

The boundary conditions for solving Eq. (10-39) are the hole concentration at the boundary between the N-region and the P-N junction ($x = 0$) and that at the outer boundary of the N-region ($x = w_n$). The first boundary condition may be written on the basis of Eq. (10-30)

$$\Delta p_n|_{x=0} = p_{n0} [\exp (eV/kT) - 1] \qquad (10\text{-}40)$$

The second boundary condition, according to Eq. (10-35), is

$$\Delta p_n|_{x=w_n} = 0 \qquad (10\text{-}41)$$

In the general form, the solution of Eq. (10-39) is

$$\Delta p\ (x) = C_1 \exp\ (x/L_p) + C_2 \exp\ (-x/L_p) \qquad (10\text{-}42)$$

Using the boundary conditions given by Eqs. (10-40) and (10-41), we may define the coefficients C_1 and C_2 as

$$C_1 = [\Delta p_n|_{x=0}/2 \sinh (w_n/L_p)] \exp (-w_n/L_p) \qquad (10\text{-}43)$$

$$C_2 = [\Delta p_n|_{x=0}/2 \sinh (w_n/L_p)] \exp (w_n/L_p) \qquad (10\text{-}44)$$

Substituting into Eq. (10-42) gives

$$\Delta p\ (x) = p_{n0} [\exp (eV/kT) - 1] \sinh [(w_n - x)/L_p] \sinh (w_n/L_p) \qquad (10\text{-}45)$$

As will be recalled, the hole diffusion current density is defined by Eq. (9-102) as

$$j_D = -eD_p\, dp/dx$$

Differentiating Eq. (10-45), we get the hole concentration gradient in our case

$$dp/dx = (p_{n0}/L_p)\,[\exp(eV/kT) - 1]\cosh[(w_n - x)/L_p]/\sinh(w_n/L_p) \tag{10-46}$$

At $x = 0$, this expression reduces to

$$dp/dx = (p_{n0}/L_p)\,[\exp(eV/kT - 1]\coth(w_n/L_p) \tag{10-47}$$

Substituting this expression into Eq. (9-102), we get the hole diffusion current density

$$j_{Dp} = (eD_p p_{n0}/L_p)\,[\exp(eV/kT) - 1]\coth(w_n/L_p) \tag{10-48}$$

A similar expression may be derived for the electron diffusion current density in the P-region

$$j_{Dn} = (eD_n n_{p0}/L_n)\,[\exp(eV/kT) - 1]\coth(w_p/L_n) \tag{10-49}$$

Multiplying Eqs. (10-48) and (10-49) by the cross-sectional area of the crystal, s, we may get expressions for the respective current components. Holes in the N-region and electrons in the P-region move in the opposite directions, but, since they carry unlike charges, the total current is equal to the sum of the electron and hole components

$$I = [(seD_p p_{n0}/L_p)\coth(w_n/L_p) + (seD_n n_{p0}/L_n)\coth(w_p/L_n)] \times [\exp(eV/kT) - 1] \tag{10-50}$$

When V is sufficiently negative, the term $\exp(eV/kT) \to 0$ and, as a consequence, a current independent of the applied voltage flows across the junction. In other words, the expression in the square brackets defines the reverse saturation current

$$I_0 = (seD_p p_{n0}/L_p)\coth(w_n/L_p) + (seD_n n_{p0}/L_n)\coth(w_p/L_n) \tag{10-51}$$

Equation (10-50) which can be written as

$$I = I_0\,[\exp(eV/kT) - 1] \tag{10-52}$$

is the *volt-ampere characteristic* of the idealized P-N junction.

As follows from Eq. (10-52), when V is sufficiently positive (forward biasing), the current across the junction increases exponentially (Fig. 10-5).

The reverse saturation current depends on the ratio w/L. If $w_n \gg L_p$ and $w_p \gg L_n$, then $\coth(w/L) =$ (approx.) 1 and, as a consequence,

$$I_0 = es(D_p p_{n0}/L_p + D_n n_{p0}/L_n) \tag{10-53}$$

If $w_n \ll L_p$ and $w_p \ll L_n$, then $\coth(w/L) =$ (approx.) L/w. Thus

$$I_0 = es(D_p p_{n0}/w_n + D_n n_{p0}/w_p) \tag{10-54}$$

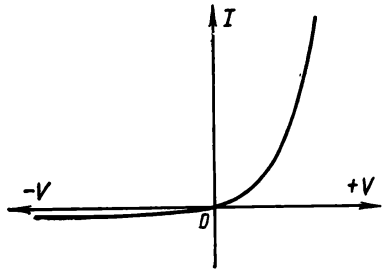


Fig. 10-5. Volt-ampere characteristic of a P-N junction

10-4. TYPES OF JUNCTIONS

In previous sections, we have discussed an idealized symmetrical P-N junction in which the P-N boundary abruptly separates two regions with different types of conduction. In other words, the region Δx in which the type and concentration of impurities change is very small ($\Delta x \to 0$). In real P-N junctions, $\Delta x > 0$, so a junction is called *abrupt* if the space-charge region (the width of the barrier region l) is much larger than the region Δx.

The Graded P-N Junction. In practice, we often use P-N junctions in which the region Δx is comparable with the barrier region. These are *graded* P-N junctions. As a rule, in making a graded P-N junction, a semiconductor wafer has impurities of the opposite types diffused into one or two faces.

For a symmetrical graded junction, Eq. (10-24) may be written as

$$l = \sqrt[3]{9\varepsilon\varepsilon_0\varphi_c/e\,(dN_a/dx + dN_d/dx)} \tag{10-55}$$

The Asymmetrical P-N Junction. This is made up of semiconductors having different impurity concentrations and, as a consequence, different concentrations of majority and minority carriers. If the difference in concentrations of majority carriers exceeds one order of magnitude, we speak of a *unipolar* P-N junction.

Figure 10-6 shows approximate plots of carrier concentrations, the distribution of potentials, field strength and concentration of immobile charges.

As $|\,dn/dx\,| < |\,dp/dx\,|$, the diffusion of holes from the P- to N-region proceeds at a higher rate than the diffusion of electrons in the reverse direction. So the forward current across the junction is mainly due to holes moving from the P- to N-region, and the reverse current is due to holes moving from the N- to P-region ($p_n > n_p$).

The condition for equilibrium is defined by Eq. (10-5) as before, but since $j_{Dp} > j_{Dn}$ and $j_{dp} > j_{dn}$, it may be re-written as

$$j_{Dp} + j_{dp} = 0 \tag{10-56}$$

In so far as $N_a \gg N_d$, the field penetrates deeper into the N-region, that is, $l_n > l_p$. In order to find the width of the barrier region in an abrupt asymmetrical junction, we may use Eq. (10-24), on putting $N_a + N_d = (\text{approx.})\, N_a$

$$l = (\text{approx.})\, l_n = (\text{approx.}) \sqrt{2\varepsilon\varepsilon_0\varphi_c/eN_d} \tag{10-57}$$

For a graded asymmetrical junction, Eq. (10-55) is re-written as

$$l = (\text{approx.}) \sqrt[3]{9\varepsilon\varepsilon_0\varphi_c/e\,(dN_d/dx)} \tag{10-58}$$

As with the symmetrical P-N junction, an applied voltage upsets the equilibrium. With forward biasing, the height of the potential barrier decreases and a forward current flows across the junction, which, in this case, is mainly due to holes moving from the P- to N-region.

Reverse biasing leads to a reverse minority-carrier saturation current which is mainly constituted by holes travelling from the N- to P-region.

Asymmetrical P-N junctions are most common in semiconductor devices. Here, a semiconductor having a high concentration of majority carriers is called the *emitter*, and the other the *base*.

P-I, N-I, P⁺-P and N⁺-N Junctions. Figure 10-7 illustrates the energy-band diagram and the distribution of concentrations in the barrier region of a P-I region. The difference in concentration at the contact of such semiconductors ($p_p > p_i$ and $n_i > n_p$) results in the diffusion of holes into the intrinsic semiconductor and electrons into the P-type semiconductor. The potential difference at the junction is established by acceptor ions in the P-type semiconductor and, in contrast to a conventional P-N junction, by holes in the intrinsic semiconductor. The barrier region mostly stretches into the I-region because its resistivity is higher.

The picture is about the same in a junction made up of a highly-alloyed and a low-alloyed semiconductor (P^+-P junction). In this case, the height of the potential barrier is smaller than it is in a P-I junction, because the difference in hole concentration is lower.

Metal-N-Type Semiconductor Junctions. Suppose that a junction is composed of a metal and an N-type semiconductor whose work function is below that of the metal ($e\varphi_{0n} < e\varphi_0$). Separate energy-band diagrams for these materials and an energy-band diagram for the junction at equilibrium are shown in Fig. 10-8. Since $e\varphi_{0n} < e\varphi_0$, electrons move from the conduction band of the N-type semiconductor into the metal, thereby charging it negatively. Near the juncti-

on, the semiconductor has a layer which is depleted of majority carriers and has an uncompensated positive charge established by donor atoms. This gives rise to an electric field $\mathscr{E}_c$ which opposes any further motion of electrons into the metal. The field sweeps

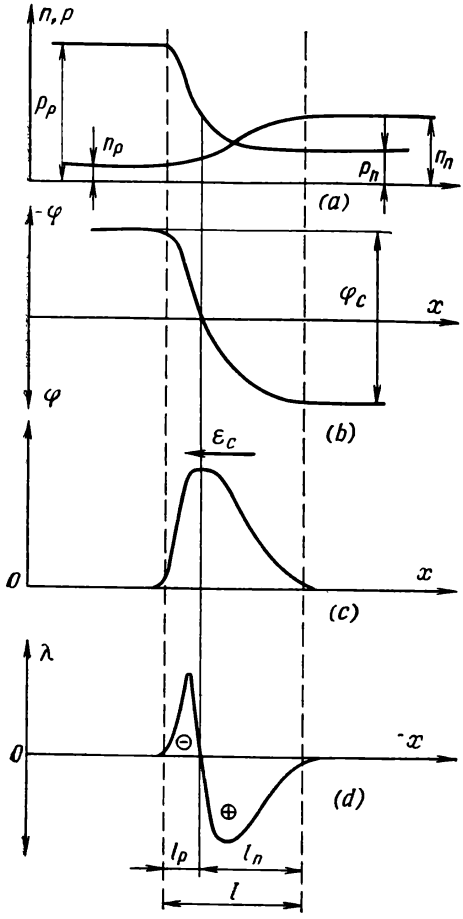


Fig. 10-6. Asymmetrical *P-N* junction
(*a*) concentration of mobile carriers; (*b*) potential distribution; (*c*) field-strength; (*d*) concentration of immobile carriers

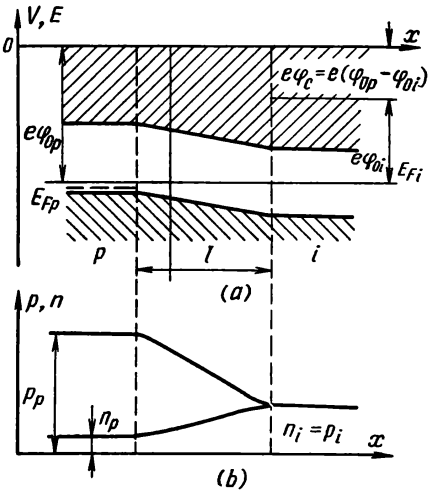


Fig. 10-7. (*a*) Energy diagram of a *P-N* junction in equilibrium; (*b*) concentration of mobile charge carriers

away free electrons (in the conduction band) and drives holes towards the junction (in the valence band). Under conditions of equilibrium, the Fermi levels in the metal and semiconductor are equalized. The barrier region formed at the junction lies mostly in the semiconductor because its resistivity is considerably higher.

With forward biasing, the potential barrier becomes lower, the resistance of the depletion layer decreases, and a current due to electrons moving into the metal flows across the junction. When the junction is reverse-biased, the potential barrier becomes higher, but the increased field allows holes to move into the metal. This current

is low because the concentration of minority carriers in the N-type semiconductor is small. Thus, this junction is also unipolar, that is, it has unidirectional conduction.

The metal-semiconductor junction is also known as the *Schottky junction.*

Metal-P-Type Semiconductor Junctions. Figure 10-9 illustrates a junction made up of a metal and P-type semiconductor whose

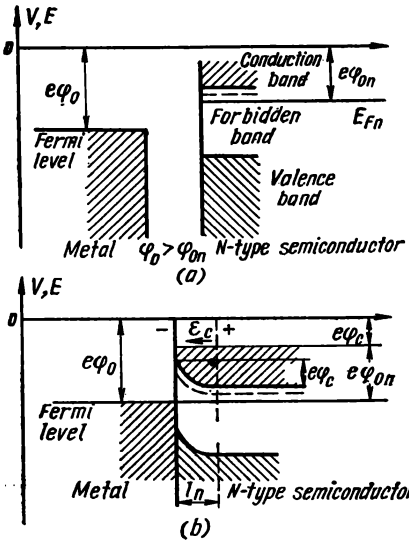


Fig. 10-8. Energy diagrams of a metal and an N-type semiconductor

(*a* out of contact; (*b*) depletion layer at the junction

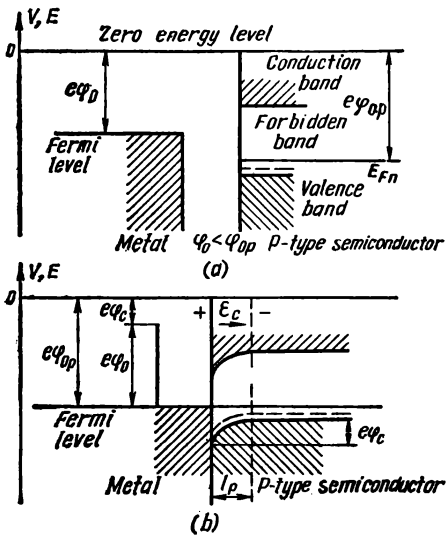


Fig. 10-9. Energy diagrams of a metal and a P-type semiconductor

(*a*) out of contact; (*b*) depletion layer at the junction

work function exceeds that of the metal ($e\varphi_{0p} > e\varphi_0$). Here, the barrier region lies in the P-type semiconductor. Electrons travelling from the metal to the semiconductor charge it negatively. The electric field at the junction opposes the electrons and drives holes out of the barrier region. This junction is unipolar, too, because forward biasing raises the electron flow from the metal, whereas reverse biasing results only in a current due to minority carriers, that is, electrons moving from the semiconductor into the metal. The density of this current is low.

Inversion Region. The processes at the junction somewhat differ if the difference in work function between the metal and the semiconductor is appreciable. Figure 10-10 shows an energy-band diagram for a metal-N-type semiconductor junction when $e\varphi_0 \gg e\varphi_{0n}$.

Because $e\varphi_c$ is considerable, the energy bands in the semiconductor bend at the edges so much that a layer with P-type conduction,

called the *inverse region*, is formed in a part of the forbidden band, l_{in}. This is borne out by the fact that the Fermi level is located below the centre of the forbidden band, which is typical of semiconductors. The inverse region is formed because the N-type semiconductor is

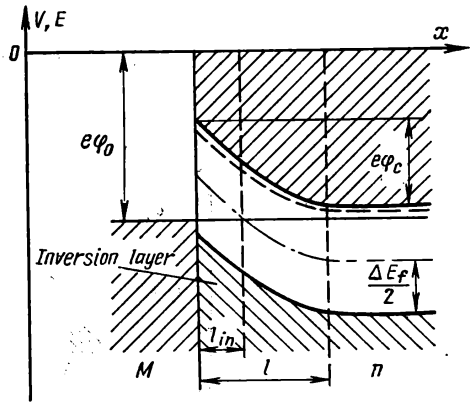


Fig. 10-10. Inversion layer at a metal-N-type-semiconductor junction

short of free electrons necessary for it to reach equilibrium. The N-type semiconductor attains equilibrium because some valence electrons move into the metal and this, of course, leaves an excess of holes near the junction. In this way a graded junction is formed near the junction.

A similar result is observed with a junction made up of a metal and a P-type semiconductor, if $e\varphi_0 \ll e\varphi_{0p}$.

The Ohmic Junction. If a junction is composed of a metal and an N-type semiconductor for which $e\varphi_0 < e\varphi_{0n}$, or of a metal and a P-type semiconductor for which $e\varphi_0 > e\varphi_{0p}$, a layer with an increased concentration of majority carriers is formed near the junction. This junction with an increased conductivity as compared with the semiconductor is called an *ohmic* junction because it has no unipolar properties. Forward or reverse biasing affects only the extent to which the layer near the junction is enriched with majority carriers.

Such junctions are used in making contact leads from the semiconductor regions forming a P-N junction.

The Heterogeneous Junction. This is a junction formed by two semiconductors differing in the width of the forbidden bands. Heterogeneous junctions are fabricated by special processes, such as the epitaxial growth from a gaseous phase or vacuum deposition of one semiconductor onto another.

Because of their difference in lattice constant various defects are formed at the boundary between the two semiconductors, and these defects may be regarded as trapping centres or centres of generation and recombination of charge carriers.

Because the two semiconductors differ in the width of the forbidden bands, the boundaries of the energy bands at the junction are broken. The energy-band diagrams of some heterogeneous junctions are shown in Fig. 10-11. The boundaries of the energy bands near the junction are bent according to whether the region at the junction is depleted of or enriched by charge carriers, whereas the distances between the band boundaries and the Fermi level depend on the concentration of boundary states. The energy gaps between the respective energy bands remain the same as they were before the contact.

When an external voltage is applied to such a junction, it loses its equilibrium, and a current flows across the junction. The nature of this current depends on several factors. For example, if the energy barrier for holes is higher than it is for electrons, the resultant current is mainly due to electrons, and vice versa. If the concentration of boundary states is high, the current across the junction is due to generation and recombination of charge carriers.

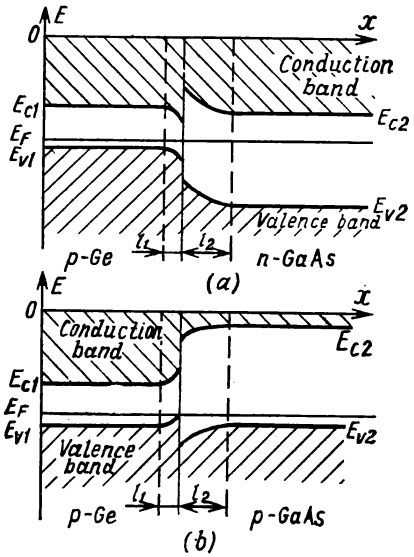


Fig. 10-11. Energy diagrams of heterogeneous junctions

(*a*) *P*-type germanium—*N*-type gallium arsenide; (*b*) *P*-type germanium—*P*-type gallium arsenide

10-5. BREAKDOWN OF THE P-N JUNCTION

At some value of the reverse voltage, a point is reached where the reverse current increases suddenly. This event is called a *breakdown*. The breakdown of a *P-N* junction may occur due to either the effect of a strong electric field in the barrier region or heat generated by the reverse current.

Avalanche Breakdown. This occurs in *P-N* junctions made up of low-alloyed semiconductors when the barrier region is sufficiently wide. If a high reverse voltage is applied to a *P-N* junction, the resultant field in the barrier region, $\mathscr{E} = \mathscr{E}_c + \mathscr{E}_{ext}$, may turn out to be so strong that the minority carriers accelerated by the junction field acquire an energy sufficient for semiconductor atoms to be ionized. On their way through the wide barrier region, holes and electrons form new pairs of mobile carriers which are in turn accelerated by the field $\mathscr{E}$ and ionize new atoms. Impact ionization in the

barrier region gives rise to an avalanche of mobile charge carriers, and the reverse current suddenly increases.

This process is described in terms of the *multiplication factor*, M, defined as the ratio of the number of carriers leaving the junction to the number, N_1, of carriers entering the junction

$$M = (N_1 + N_2 + N_2')/N_1 \tag{10-59}$$

where N_2 is the number of particles produced by impact ionization by electrons, and N_2' is the number of particles produced by ionization by holes.

The multiplication factor, M, naturally depends on the applied reverse voltage, the resistivity of the contacting semiconductors (their degree of alloying), and some other factors. The reverse voltage at which the multiplication factor, M, tends to infinity is called the *avalanche breakdown voltage*, designated $V_{bd,\,av}$. The multiplication factor is related to this voltage by the following formula

$$M = 1/[1 - (V/V_{bd,\,av})^b] \tag{10-60}$$

where $b = 3$ for P-type silicon and N-type germanium, and 5 for P-type germanium and N-type silicon.

The breakdown voltage can be found by the following empirical formula

$$V_{bd,\,av} = A\rho_b^m \tag{10-61}$$

Here, ρ_b is the resistivity of the base in ohms cm; m = (approx.) 0.6 for germanium and 0.7 for silicon; A = 83 for N-type germanium, 52 for P-type germanium, 86 for N-type silicon and 23 for P-type silicon.

Avalanche breakdown is characterized by a sudden increase in reverse current with the reverse voltage remaining practically unchanged (Fig. 10-12).

Zener Breakdown. This occurs in junctions composed of semiconductors having lower resistivities and narrow forbidden bands. As has been shown in Sec. 9-3, a strong applied electric field causes the boundaries of the energy bands to shift and a sufficiently thin potential barrier to form near the junction. The probability that particles will tunnel through this barrier is defined by Eq. (9-8)

$$P = (\text{approx.}) \exp [-(4\pi/eh\mathscr{E}) \sqrt{2m^* (\Delta E_f)^3}]$$

The current due to the tunneling of particles is called the *tunnel current*, I_t.

Zener breakdown is said to occur when the tunnel current is ten times the reverse current, I_0. For germanium, Zener breakdown happens at a field strength of $\mathscr{E}_{bd,z}$ = (approx.) 2×10^5 V/cm, and for silicon at $\mathscr{E}_{bd,z}$ = (approx.) 4×10^5 V/cm.

The field strength $\mathscr{E}_{bd,\,z}$ depends on both the applied voltage, V, and the resistivity of the semiconductors. The breakdown voltage, $V_{bd,\,z}$, can be found by the following empirical expressions
— for germanium

$$V_{bd,\,z} = \text{(approx.)}\ 19 \times 10^3 \rho_n + 94 \times 10^2 \rho_p \qquad (10\text{-}62)$$

— for silicon

$$V_{bd,\,z} = \text{(approx.)}\ 20 \times 10^3 \rho_n + 73 \times 10^2 \rho_p \qquad (10\text{-}63)$$

where ρ is the resistivity in ohms m.

In high-resistance semiconductors, $V_{bd,\,z} > V_{bd,\,av}$; in low-resistance semiconductors, Zener breakdown takes place at lower voltages, $V_{bd,\,z} < V_{bd,\,av}$.

The reverse current characteristic under conditions of Zener breakdown is the same as under conditions of avalanche breakdown (Fig. 10-12).

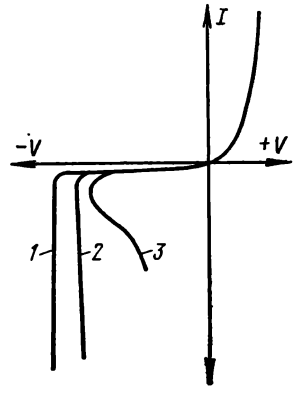


Fig. 10-12. Breakdown of the *P-N* junction
(*a*) avalanche breakdown; (*b*) Zener breakdown; (*c*) thermal breakdown

Thermal Breakdown. This occurs due to the heating of a *P-N* junction by a high reverse current flowing across it. If the amount of heat generated in the *P-N* junction exceeds the amount abstracted, the temperature of the junction increases. The number of minority carriers and the reverse current rise, the junction is heated still more, and so on. The thermal breakdown voltage, $V_{bd,\,th}$, depends on the reverse current, resistivity of the semiconductor, heat abstraction, ambient temperature and some other factors. Under conditions of thermal breakdown, the relation $I_0 = f(V)$ has a negative differential resistance region which is a distinction of this curve (Fig. 10-12).

10-6. CAPACITANCES OF THE P-N JUNCTION

When discussing the physical processes in the *P-N* junction, we showed that the ionized impurity atoms (Fig. 10-2*d*) and the mobile charge carriers penetrating into the barrier region produce space charges on either side of the *P-N* boundary.

In addition, an external voltage applied to the junction causes injection and extraction of minority carriers which affect the concentrations of majority carriers in the semiconductor regions near the depletion layer (Figs. 10-3*c* and 10-4*c*).

The *P*- and *N*-regions on either side of the *P-N* boundary carrying unlike mobile charges may be treated as the plates of an equivalent

capacitor connected in parallel with the junction. This apparent capacitance of the junction is called the *junction capacitance.*

The space charges at and near the junction vary in accordance with the applied voltage, because it affects the width of the barrier region l, and also the concentration of majority and minority carriers near the boundaries of the junction. So, in the general case, $C_j = f(V)$.

The space charges within and beyond the barrier region depend on the applied voltage differently. Accordingly, we customarily assume that the junction capacitance has two components.

One of them, the *barrier capacitance*, is associated with the redistribution of charges in the barrier region. The other, known as the *diffusion capacitance*, is associated with the redistribution at the junction boundaries, caused by injection and extraction of carriers.

Barrier Capacitance. Let us discuss this capacitance in an abrupt asymmetrical P-N junction in which $N_a \gg N_d$. In this case, most of the depletion layer lies in the N-type semiconductor. The width of this layer is defined by Eq. (10-57)

$$l = (\text{approx.})\ l_n = (\text{approx.}) \sqrt{2\varepsilon\varepsilon_0\varphi_c/eN_d}$$

Under conditions of reverse biasing, the barrier region becomes wider

$$l'' = (\text{approx.})\ l_n'' = (\text{approx.}) \sqrt{2\varepsilon\varepsilon_0 (\varphi_c + |V|)/eN_d}$$
$$= (\text{approx.})\, l \sqrt{(\varphi_c + |V|)/\varphi_c} \qquad (10\text{-}64)$$

If we imagine the barrier capacitance as a capacitor with charges $Q_p = Q_n$ on its plates, we may write, for example,

$$Q_n = eN_d sl \qquad (10\text{-}65)$$

where s is the cross-sectional area of the P-N junction. The charge Q_n is not proportional to the voltage, so $C_b = dQ/dV$. Substituting Eq. (10-64) into Eq. (10-65) and differentiating the latter with respect to the voltage, we get the differential barrier capacitance

$$C_b = (\varepsilon\varepsilon_0 s/l) \sqrt{\varphi_c/(\varphi_c + |V|)} \qquad (10\text{-}66)$$

As the reverse voltage rises, the barrier capacitance of the abrupt asymmetrical junction varies approximately in proportion to the square root of the voltage.

The function $C_b = f(V)$ depends to a great extent on the distribution of the impurity concentration in the junction region. For example, for a graded junction, C_b varies approximately in inverse proportion to the cubic root of the voltage. The $C_b = f(v)$ curve is the steepest in junctions where variations in the impurity concen-

tration are described by a complex function which is linear with x in one region, and represented by a hyperbolic function in the other.

The relation $C_b = f(V)$ is called the *volt-farad characteristic* of the P-N junction (Fig. 10-13). In this figure, C_{b0} is the barrier capacitance at $V = 0$.

When a junction is forward-biased, the barrier region becomes narrower and the barrier capacitance increases. Under the circumstances, however, an increase in charges due to the injection of minority carriers has a considerably greater effect on C_j than a decrease in the barrier region width l. In other words, under conditions of forward biasing, the junction capacitance is mainly determined by the diffusion capacitance C_D.

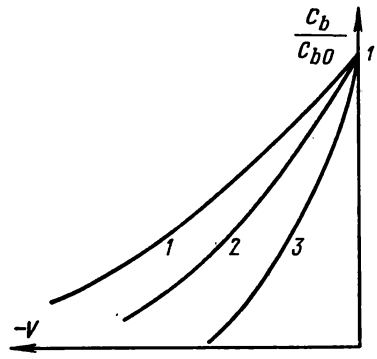


Fig. 10-13. Volt-farad characteristics of P-N junctions

1—graded junction; *2*—abrupt junction; *3*—junction with a complex impurity concentration function

Diffusion Capacitance. As already noted, this capacitance depends on variations in charges at the junction boundaries caused by variations in the applied voltage V. When a junction is reverse-biased, the capacitance C_D is low because the concentration of minority carriers varies from the equilibrium value to near zero only when V varies from zero to a small negative value (see Fig. 10-4*c*). As the reverse voltage rises further, the distribution of minority carriers remains practically unchanged.

Under conditions of forward biasing, there is a considerable change in space charge in the semiconductor regions at the junction boundaries. Due to injection of minority carriers, especially at $\delta \geqslant 1$, an increase in forward voltage may lead to a considerable rise in this charge.

Let us find the increment, ΔQ, in hole charge in the N-region by integrating the function $p(x)$ defined by Eq. (10-35) over the entire region from 0 to w_n

$$\Delta Q = es \int_0^{w_n} \Delta p(x)\, dx = [esL_p \Delta p|_{x=0}/\tanh(w_n/L_p)]\,[1 - \operatorname{sech}(w_n/L_p)] \tag{10-67}$$

Using Eqs. (9-116) and (10-51) we get for the differential capacitance

$$C_D = (e\tau_p I/kT)\,[1 - \operatorname{sech}(w_n/L_p)] \tag{10-68}$$

where I is the current across the junction under conditions of forward biasing.

When $w_n > L_p$, the function sech (w_n/L_p) is approximately equal to zero and Eq. (10-68) takes the following form

$$C_D = \text{(approx.)}\ eI\tau_p/kT \tag{10-69}$$

If $w_n < L_p$, then $\operatorname{sech}(w_n/L_p) = \text{(approx.)}\ 1 - w^2/2L_p^2$ and, recalling that $L_p = \sqrt{D_p\tau_p}$, we may write

$$C_D = \text{(approx.)}\,(eI/kT)\,(w^2/2D) \tag{10-70}$$

As is seen, the diffusion capacitance is directly proportional to the forward current across the junction. When the current is sufficiently high, C_D may be several orders of magnitude higher than C_b.

Chapter Eleven

CRYSTAL DIODES

11-1. DESIGN AND CLASSIFICATION OF CRYSTAL DIODES

Definition. A *crystal diode* is a rectifying semiconductor device having one *P-N* transition region and two terminals.

Design. As its name implies, the crystal diode is a semiconductor crystal containing a *P-N* transition region fabricated by one of the above methods. To two semiconductor regions forming the transition region are welded or soldered metal-wire leads, and the whole system is enclosed in a glass, metal, plastic or ceramic case or encapsulated in a special resin. As an example, Fig. 11-1 shows an alloy silicon diode in a glass case and an unpackaged diffused diode.

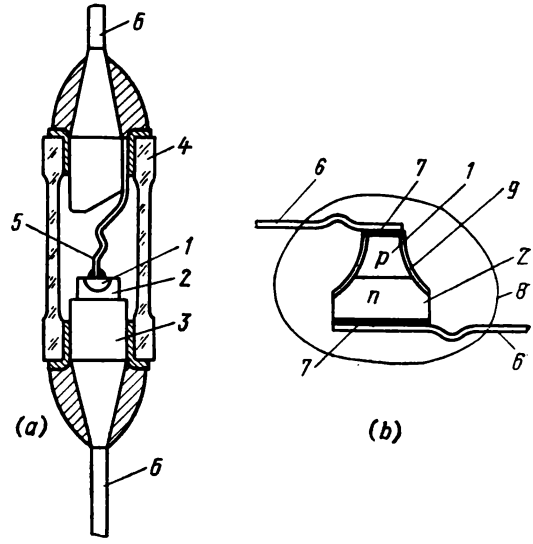


Fig. 11-1. (*a*) Alloy silicon diode and (*b*) unpackaged diffused diode

1—P-type silicon; *2*—*N*-type silicon; *3*—support; *4*—glass case; *5*—aluminium; *6*—lead; *7*—ohmic contact; *8*—resin; *9*—film of SiO_2

One of the semiconductor regions making up the *P-N* transition region has an increased concentration of impurities and serves as the *emitter*, and the other operates as the *base*.

Most crystal diodes utilize asymmetrical *N-P* and *P-N* transition regions. Use is also made of metal-semiconductor, *P-I*, *N-I* and some other junctions.

Classification. Crystal diodes are many and diverse and may be classed in a variety of ways. To begin with, they may be classed according to the method of manufacture and the type of *P-N* transition used into the *junction* type and the *point-contact* type. The former are usually fabricated by the alloy or diffusion process. In point-contact diodes the surface area of the transition region is much smaller than in junction diodes. They are manufactured by welding a pointed wire (called a *catwhisker*) to a diode base and adding simultaneously an alloying material.

Crystal diodes may also be classified according to the host material into *germanium*, *silicon*, *gallium-arsenide* and some other types. One of the main characteristics, however, is the purpose of the diode. Accordingly, crystal diodes are divided into *rectifying* diodes, *detector* diodes, *conversion* diodes, *switching* diodes, *voltage* stabilizer diodes (*Zener* diodes), *parametric* diodes, etc. In accordance with the main physical process utilized, they may be classed into *avalanche* diodes, *tunnel* diodes, *photodiodes*, *light-emitting* diodes, and so on.

In this chapter, we shall discuss most types, except photodiodes and light-emitting diodes which are described in Chapter 14.

Crystal diodes used in the microwave region (such as parametric, switching, avalanche and Gunn diodes) are discussed in a course on microwave and quantum devices.

In the Soviet Union, the diagram symbols for crystal diodes are set up by USSR State Standard GOST 10862-72 and GOST 2.728-73. The nomenclature for them is established by GOST 20004-74, GOST 20005-74 and GOST 18994-73.

11-2. PHYSICAL PROCESSES IN THE CRYSTAL DIODE

As already noted, a crystal diode has one *P-N* junction. So, the physical processes in the *P-N* junction discussed in Chapter 10 are, in fact, the physical processes in an idealized diode. Consequently, the volt-ampere characteristic of an idealized crystal diode is described by Eq. (10-52) derived for the *P-N* junction

$$I = I_0 [\exp (eV/kT) - 1]$$

When discussing an idealized *P-N* junction, however, we did not take into consideration some physical processes taking place in real diodes. So the relation defined by Eq. (10-52) is affected by some additional factors. Let us first discuss the main of these factors using, as an example, a diode with an *N*-type base.

Reverse Biasing. As follows from Chapter 10, the barrier region with reverse bias becomes wider, the potential barrier grows higher, so the junction and, as a consequence, the diode, offers an increased resistance. For diodes, the reverse current flowing across the junction is defined as the total conduction current flowing in the reverse direction. The reverse current is important not only in the case of reverse biasing when at $V < 0$

$$I = (\text{approx.}) - I_0 \tag{11-1}$$

but also under conditions of forward biasing. In the latter case ($V > 0$), in accordance with Eq. (10-52), we have

$$I = (\text{approx.})\, I_0 \exp(eV/kT) \tag{11-2}$$

and, as a consequence, the forward volt-ampere characteristic depends on I_0 as well.

In deriving the volt-ampere characteristic of an idealized *P-N* junction in Sec. 10-3 [see Eq. (10-52)], we assumed that the reverse current was only due to minority carriers which, in semiconductors, are mainly produced by thermal electron-hole generation. Accordingly, this current is called *thermal current.* In real devices, the thermal current is only a part of the reverse current which contains some other components such as current due to carrier generation at the junction and leakage current.

Let us first discuss the factors affecting thermal current.

Thermal Current. Assuming that $w_n > L_p$ and $w_p > L_n$, thermal current is defined by Eq. (10-53)

$$I_0 = es\,(D_p p_{n0}/L_p + D_n n_{p0}/L_n)$$

where p_{n0} and n_{p0} are the equilibrium concentrations of minority carriers and s is the junction area.

When the emitter and base are thin, that is, $w_n \ll L_p$ and $w_p \ll \ll L_n$, thermal current is defined by Eq. (10-54)

$$I_0 = es\,(D_p p_{n0}/w_n + D_n n_{p0}/w_p)$$

Substituting into Eq. (10-53) the expressions for the diffusion constants $D_p = L_p^2/\tau_p$ and $D_n = L_n^2/\tau_n$ derived from Eq. (9-116) gives

$$I_0 = es\,(L_p p_{n0}/\tau_p + L_n n_{p0}/\tau_n) \tag{11-3}$$

According to Eq. (9-16), the ratios p_{n0}/τ_p and n_{p0}/τ_n in the above expression are nothing but the generation rates for holes and electrons, respectively. Thus, the thermal current in an idealized *P-N* junction whose width is $l \to 0$ is due to the generation of minority carriers in the semiconductor volumes sL_p and sL_n bordering upon the *P-N* boundary.

Because of concentration gradients existing in these volumes (see Fig. 10-4*c*), minority carriers diffuse to, and the field $\mathscr{E}_c$ sweeps them across, the boundary.

As follows from comparison of Eqs. (10-53) and (10-54), at $w_n \ll \ll L_p$ and $w_p \ll L_n$, I_0 is greater than in the former case. Physically, this change in thermal current as the semiconductor regions become narrow can be explained as follows. The generation of minority carriers may take place throughout the volume, including the regions where the metal leads (ohmic contacts) are connected to the electrodes. As a rule, in these regions, the generation is most intensive. When $w > L$, the minority carriers may fail to reach the junction, so they will not take part in the motion across the barrier region. Besides, as is seen from Fig. 10-4, the concentration gradient of minority carriers at $w > L$ is smaller than at $w \ll L$, and, as a consequence, the diffusion current density due to minority carriers moving from the base towards the *P-N* junction is lower than in the latter case.

In addition, the thermal current depends on the junction area, s; an increase in s brings about an increase in I_0.

Also important is the dependence of thermal current on the concentration of minority carriers, p_{n0} and n_{p0}. If a diode has an asymmetrical *P-N* junction, and the degree of alloying of the *P*-emitter is considerably higher than that of the *N*-base ($N_a \gg N_d$), the concentration of minority carriers in the base will be higher than in the emitter, so the thermal current will mainly be constituted by holes which are the minority carriers of the base. In this case (at $w_n \ll \ll L_p$), the thermal current may be defined as

$$I_0 = (\text{approx.})\, esD_p p_{n0}/w_n \qquad (11\text{-}4)$$

As has been shown in Chapter 9, the concentration of minority carriers is defined by Eq. (9-67)

$$p_n = n_i^2/N_d = (N_c N_v/N_d) \exp(-\Delta E_f/kT)$$

In our case, $p_n = p_{n0}$. Substituting this into Eq. (11-4) gives

$$I_0 = (\text{approx.})\, esD_p n_i^2/w_n N_d \qquad (11\text{-}5)$$

Thus, the thermal current is proportional to n_i^2 and, as is seen from Eq. (9-67), significantly depends on temperature. For example, at T_0 = (approx.) 300 K, n_i = (approx.) 2.5×10^{13} cm^{-3} for germanium, and 2×10^{10} cm^{-3} for silicon. The diffusion constant D_p for germanium is approximately three and a half times that for silicon. Substituting these values in Eq. (11-5), we can see that at equal s, w_n and N_d, the thermal current in germanium at room temperature is by about six orders of magnitude higher than it is in silicon ($I_{0\text{Si}} \approx 10^{-6}\, I_{0\text{Ge}}$).

For any semiconductor, the relation between thermal current and temperature may be defined as

$$I_0(T) = I_0(T_0) \exp(\alpha \Delta T) \qquad (11\text{-}6)$$

where $\Delta T = T - T_0$, $T_0 = 300$ K, and $\alpha = 0.09$ K^{-1} for germanium and 0.13 K^{-1} for silicon.

Generation Current. In real diodes, the width of the barrier region l is finite, and the generation and recombination of charge carriers take place within and outside this region. Under the action of the electric field $\mathscr{E}_c$, the charge carriers generated at the junction leave the barrier region, thereby constituting the generation current I_g. The recombination of carriers in the barrier region results in the recombination current I_r.

The generation of carriers at the junction occurs due to the breaking of valence bonds and also ionization of impurity centres with deep-lying energy levels close to the centre of the forbidden band (Sec. 9-2).

The recombination of carriers results from the fact that carriers having an energy insufficient to overcome the potential barrier, but penetrating into the barrier region to some depth, lose their velocity in the junction field, and are driven by this field out of the region. Since such carriers stay in the barrier region for a long time, the probability that they will recombine through traps and other structure imperfections increases, and this gives rise to the recombination current I_r.

Under conditions of dynamic equilibrium ($V = 0$), the opposing streams of carriers cancel each other, so that the generation current is equal to the recombination current.

When a reverse voltage is applied to the diode ($V < 0$), the potential barrier at the junction increases, the depth to which carriers penetrate into the barrier region decreases, the probability that they will recombine at the junction goes down, and so does the recombination current. In addition, as the reverse voltage increases, the barrier region becomes wider and the volume in which the generation of mobile carriers takes place grows.

By analogy with Eq. (11-3), the generation current may be defined as

$$I_g = esln_i/(\tau_n + \tau_p) \tag{11-7}$$

The point is that the generation of carriers which defines I_g takes place in the volume sl of the barrier region where, at $T =$ (approx.) 300 K, all of the impurity atoms are ionized and the generation rate of both electrons and holes depends on their lifetimes. When the reverse voltage exceeds a few hundredths of a volt, the generation current is much higher than the recombination current, $I_g \gg I_r$.

As follows from Eq. (11-7), the difference in intrinsic carrier concentration between germanium and silicon (2.5×10^{13} cm^{-3} and 2×10^{10} cm^{-3}, respectively) results in that I_g in germanium is about one thousand times the value it has in silicon for geometrically equal junctions.

As is seen from Eq. (11-7), I_g is directly proportional to the width of the barrier region l. Because of this, I_g increases with increasing reverse voltage in the same manner as l, that is, in proportion to $|V|^{1/2}$ [see Eq. (10-37)].

Let us compare I_g with the thermal current in the diode, assuming that $w_n \gg L_p$ and that I_0 is mainly constituted by holes moving from the base to the emitter, which is true when the impurity concentration in the emitter is considerably higher than it is in the base ($N_a \gg N_d$). Then, on the basis of Eq. (11-3), we have

$$I_0 = (\text{approx.})\, esL_p p_{n0}/\tau_p \tag{11-8}$$

Using Eq. (9-67), we get

$$I_0 = (\text{approx.})\, esL_p n_i^2/\tau_p N_d \tag{11-9}$$

Putting $\tau_n = (\text{approx.})\, \tau_p$ and recalling Eq. (11-7), we may write the current ratio as

$$I_g/I_0 = lN_d/2L_p n_i \tag{11-10}$$

The difference in this ratio between germanium and silicon diodes (at the same N_d) is mainly attributed to the difference in n_i for these materials, because L_p and l (due to the difference in permittivity) for germanium and silicon differ less significantly (see Table 9-1). The solution of Eq. (11-10) shows that, at $N_d = (\text{approx.})$ 10^{15} cm^{-3} and $T = (\text{approx.})$ 300 K, the generation current in the germanium diode is several tenths of I_0, and in the silicon diode, approximately $10^3\, I_0$.

To define the relation $I_g = f(T)$, we may use Eq. (11-6) with α reduced by one half.

Leakage Current. Surface phenomena give rise to what is known as the *leakage current*, I_l, which is comparable with I_0 in germanium and I_g in silicon diodes at very negative voltages. Leakage current increases in proportion to reverse voltage. It may significantly vary with time and ambient temperature, thereby affecting the stability of the downward slope of the volt-ampere characteristic, because I_l accounts for a considerable part of the reverse current.

Total Reverse Current. As is seen from the foregoing, the reverse current in the diode, I_{rev}, is the sum of several components

$$I_{rev} = I_0 + I_g + I_l \tag{11-11}$$

At $T = (\text{approx.})$ 300 K , the reverse current in germanium diodes is mainly constituted by the thermal current, and in silicon diodes, by the generation current. Since I_0 is more sensitive to temperature variations, at temperatures over 100°C the thermal current prevails in silicon diodes.

As it follows from the above relations, the reverse current in germanium diodes is generally by several orders of magnitude higher than it is in silicon diodes.

Forward Biasing. With forward biasing, the potential barrier at the junction becomes lower, the state of equilibrium is disturbed, injection of charge carriers begins, and the junction passes the forward current, I_f, which is defined as the current flowing in the direction in which the diode offers a low resistance.

In contrast to an idealized junction, the forward current and its relation to the applied forward voltage in a real diode depend on some physical processes taking place both in the junction itself and in the base region. Let us discuss the main processes.

Recombination Current. Under conditions of forward biasing when the potential barrier at the junction becomes lower and the number of carriers penetrating into the barrier region and the time during which they stay in this region increase, the recombination of electron-hole pairs prevails over their generation, and the recombination current is higher than the generation current, $I_r > I_g$.

I_r is defined by an approximate equation given in reference [24]

$$I_r = esl\,[n_i/(\tau_p + \tau_n)] \exp(eV/2kT) \qquad (11\text{-}12)$$

As is seen, the recombination current rises exponentially with rising voltage V in much the same manner as the diffusion current defined by Eq. (11-2). For I_r, however, this relation is less pronounced because, in contrast to Eq. (11-2), the exponent has a factor $^1/_2$.

In so far, as according to Eq. (11-12), I_r depends on the concentration n_i, its value is different for silicon and germanium diodes. The contribution of this current to the forward current in a diode may be estimated by comparing I_r with the thermal current. Similarly to the generation current, I_r in germanium diodes is a small fraction of the thermal current I_0; in silicon diodes, $I_r > I_0$, so at low forward voltages the recombination current constitutes a considerable part of the forward current; as V increases, the diffusion current increases more rapidly and begins to prevail in the forward current.

Injection of Charge Carriers. Since in real diodes the degree of alloying of the emitter is generally much higher than that of the base

$$N_a \gg N_d \qquad (11\text{-}13)$$

the forward current is mainly due to the injection of holes from the emitter into the base. The current due to the injection of minority carriers into the base may be defined as a part of the total current across the junction, in terms of the injection ratio

$$\gamma = I_p/(I_p + I_n) \qquad (11\text{-}14)$$

where I_p and I_n are the hole and electron components of the current across the junction at the P-N boundary.

The factor γ may be defined by means of Eqs. (10-48) and (10-49) where the constants D_p and D_n are expressed in terms of the mobilities

according to Eqs. (9-101) and (9-103) and the concentrations p_{n0} and n_{p0} are replaced by p_{p0} and n_{n0}, respectively. Now, Eq. (11-14) for a diode with extended emitter and base regions may be written as

$$\gamma = \text{(approx.) } 1 - \rho_e/\rho_b \tag{11-15}$$

As a rule, it is the relation (11-13) that applies to diodes, and, as a consequence, γ is only slightly below unity, being between 0.9 and 0.99. For example, at $N_a = 10^{18}$ cm^{-3} ($\rho \approx 0.07$ ohms·cm) and $N_d = 10^{15}$ cm^{-3} ($\rho \approx 5$ ohms·cm), γ = (approx.) 0.986.

A second important parameter characterizing the injection of carriers is the injection level, δ, which has been defined by Eq. (10-32) in Sec. 10-3. This factor defines the nature of the processes taking place in the diode base.

Physical Processes in the Base. The injection of charge carriers gives rise to excess of minority carriers $\Delta p = p_n - p_{n0}$ in the base next to the barrier region (see Fig. 10-3*c*).

Variations in the concentration of minority carriers in the base due to their diffusion away from the boundary of the barrier region are described by Eq. (10-33) for a thick base ($w_n > L_p$)

$$p(x) = p_{n0} + \Delta p \exp(-x/L_p)$$

and by Eq. (10-35) for a thin base ($w_n \ll L_p$)

$$p(x) = p_{n0} + \Delta p\,(1 - x/w_n)$$

Base Spreading Resistance. When the injection level is low, the charge carriers injected into the base have a small effect on the concentration of charge carriers in the base, and the base spreading resistance depends on its geometrical dimensions and resistivity, ρ_b, that is, the impurity concentration

$$r_{b0} = \rho_b w_n/s \tag{11-16}$$

where s is the cross-sectional area of the base.

When the base is low-alloyed, its spreading resistance under conditions of forward biasing may be comparable with the resistance of the junction. In the circumstances, the external voltage cannot be treated as applied entirely to the junction, as was assumed in the analysis of the idealized *P-N* junction. In addition it is necessary to take into account the voltage drop across the base spreading resistance

$$V = V_j + V_b \tag{11-17}$$

Modulation of Base Resistance. When the injection level is high, the contribution of the holes injected into the base to the total concentration of mobile charge carriers becomes rather significant, and the base spreading resistance may markedly decrease. This is known as the *base resistance modulation.* For a thin base ($w_n \ll L_p$),

the base spreading resistance as a function of the injection level δ may be defined by an approximate expression given in reference [24]

$$r_b(\delta) = (r_{b0} \log_e \delta)/\delta \qquad (11\text{-}18)$$

When δ approximately ranges between 15 and 20, the base resistance may decrease to a fraction of r_{b0}.

The Field in the Diode Base. As is shown in Sec. 10-1, the injection of charge carriers into the base results in excess holes near the boundary of the barrier region. This gives rise to an internal electric field in the base, whose strength vector, $\mathscr{E}_b$, is directed from the junction to the base. Under the action of this field, electrons move from the bulk of the base toward the *P-N* junction where they form a space charge approximately equal to the hole charge.

This, however, cannot fully neutralize the positive hole charge, because otherwise, the field $\mathscr{E}_b$ attracting electrons from the bulk of the base would vanish. There is a certain residual field in the base, which depends on the excess hole concentration, Δp, over the concentration of majority carriers, n_n, or, which is the same, the injection level δ.

When the injection level is low ($\Delta p \ll n_n$) the field is weak, and the holes injected into the base move from the *P-N* junction into the bulk of the base only because of their concentration gradient. The effect of the field $\mathscr{E}_b$ on this motion is negligibly small. As insignificant is the effect that the space charge exerts on the charge neutrality. Since this space charge is relatively small, the base remains quasi-neutral.

This is not exactly the case when the injection level is high ($\Delta p > > n_n$). Under the circumstances, the field $\mathscr{E}_b$ is strong enough to cause a significant drift of minority carriers—holes—in the base. In so far as the hole drift adds to their diffusion, the velocity of the holes moving from the junction to the base increases.

The existing field $\mathscr{E}_b$ can easily be determined by Eqs. (10-7) and (10-8) describing hole and electron current densities.

In our case, we may write

$$j_n = en_n\mu_n\mathscr{E}_b + eD_n\,\partial n/\partial x \qquad (11\text{-}19)$$

$$j_p = ep_n\mu_p\mathscr{E}_b - eD_p\,\partial p/\partial x \qquad (11\text{-}20)$$

Adding together these equations (on putting $\partial p/\partial x = \partial n/\partial x$) and solving them for $\mathscr{E}_b$, we get

$$\mathscr{E}_b = j/e\,(\mu_n n_n + \mu_p p_n) - [(D_n - D_p)/(\mu_n n_n - \mu_p p_n)]\,\partial p/\partial x \qquad (11\text{-}21)$$

The first term in this equation is the field in the base occurring due to the voltage drop produced across the base spreading resistance by the current flowing through the diode; the second term is the field developed due to the concentration gradient of majority and minority carriers.

11-3. THE VOLT-AMPERE CHARACTERISTIC OF THE REAL DIODE

The physical processes in the real diode discussed in the previous section make it possible to define the main differences between the characteristic of the real device and that of the idealized diode described by Eq. (10-52).

The Downward Slope of the Volt-Ampere Characteristic. As already noted, the forward current in any diode is the sum of several components whose magnitudes depend on the material. Figure 11-2

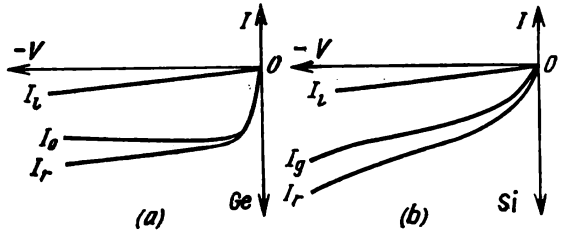


Fig. 11-2. Reverse current components
(*a*) in a germanium diode; (*b*) in a silicon diode

shows the main components of the reverse current in silicon and germanium diodes. The plots of Figs. 11-2*a* and 11-2*b* are drawn to different scales because I_0 in a germanium diode is several orders of magnitude higher than in a silicon diode.

In Sec. 10-5, we have discussed different mechanisms by which a breakdown occurs in a *P-N* junction. The volt-ampere characteristics under conditions of the avalanche, Zener and thermal breakdowns have been shown in Fig. 10-12.

The Upward Slope of the Volt-Ampere Characteristic. According to Eq. (10-52), the current through a diode should vary exponentially with the forward voltage. The real characteristics, however, differ from this idealized law because of some factors discussed in Sec. 11-2.

Since the forward current strongly depends on the applied voltage, the upward slope of the volt-ampere characteristic is usually described by a relation in which I is the argument. This relation can easily be derived by taking a logarithm of Eq. (10-52) and resolving the result for V

$$V = (kT/e) \log_e (I/I_r + 1) \qquad (11\text{-}22)$$

As is seen, the voltage V corresponding to a specified forward current I is inversely proportional to the reverse current, I_r. For silicon diodes whose I_r is much smaller than that of germanium diodes, the initial portion of the upward slope is very gradual (Fig. 11-3*a*).

In silicon diodes, the forward current within this portion is mainly due to the recombination of carriers at the junction, which at $V > 0$ prevails over thermal generation.

The volt-ampere characteristic also depends on the junction area, s (Fig. 11-3*b*). As s increases, the thermal current rises and, as a

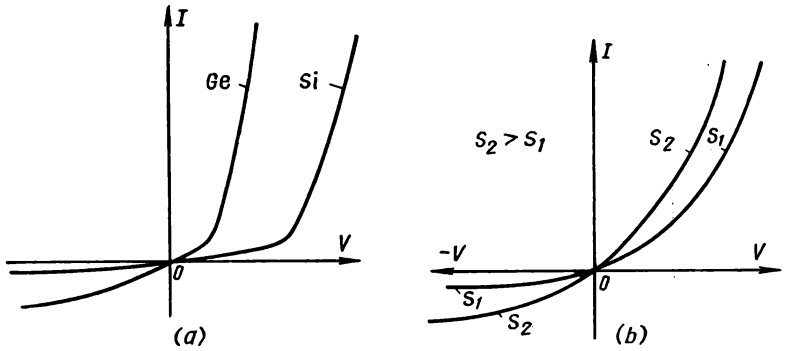


Fig. 11-3. (*a*) Volt-ampere characteristics of silicon and germanium diodes and (*b*) volt-ampere characteristic of a diode as a function of the junction area

consequence, the upward slope of the characteristic runs more steeply.

Also, the relation $V = \varphi(I)$ is markedly affected by the base resistance, r_b. As already noted, the forward current flowing through the base gives rise to a voltage drop

$$V_b = Ir_b = I\rho_b w/s \qquad (11\text{-}23)$$

This voltage drop is rather high already at $I \geqslant 2$ mA. Recalling Eq. (11-23), the expression for the volt-ampere characteristic may be written as

$$I = I_r\{\exp[e(V - Ir_b)/kT] - 1\} \qquad (11\text{-}24$$

or

$$V = (kT/e)\log_e(I/I_r + 1) + Ir_b \qquad (11\text{-}25)$$

As a result, the portion of the volt-ampere characteristic corresponding to $I \geqslant 2$ mA, or, in high-power diodes, $I \geqslant 20$ mA, becomes more gradual (Fig. 11-4). This part of the characteristic is often called *ohmic*. When the injection level is high and the number of carriers injected into the base is great, the base resistance somewhat decreases and the volt-ampere characteristic runs more steeply.

For any real diode, the initial portion of the upward slope of the volt-ampere characteristic differs from the curve described by Eq. (10-52). In germanium diodes, its slope depends on the thermal cur-

rent, and in silicon diodes, on the recombination current. In germanium diodes, a sudden rise in the forward current generally takes place at lower values of the forward voltage.

The beginning of the sharply rising portion is close to an exponential curve; here, the main contribution to the current is made by the diffusion of the carriers injected into the base (a low injection level). As the current rises further, the effect of the base spreading resistance and other processes typical of a relatively high injection level become more pronounced. The boundaries between the various portions of the volt-ampere characteristic, arbitrary as they are, and the relation $I = \varphi(V)$ itself considerably differ for germanium and silicon diodes, thick-base and thin-base diodes, etc. For example, for silicon diodes, the transition from the initial gradual portion to the high-injection-level region is more pronounced than for germanium diodes; in germanium diodes, the ohmic portion is steeper because the carriers mobility in germanium is much higher than in silicon, and so on.

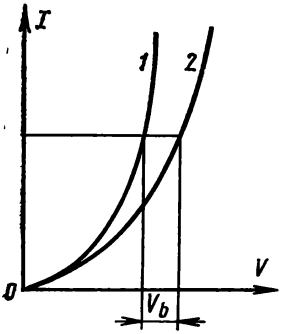


Fig. 11-4. Upward (forward) slope of the diode volt-ampere characteristic

1—exponential; *2*—modified by the voltage drop across the base spreading resistance

The Effect of Temperature Variations on the Volt-Ampere Characteristic. The effect of temperature on the performance of semiconductor devices, in particular, crystal diodes, is considerable.

Figure 11-5*a* and *b* shows the volt-ampere characteristics of the Soviet-made type Д9И germanium diode and the Soviet-made type Д106А silicon diode. As is seen, temperature variations affect both the downward and upward slopes of the characteristic.

The thermal current I_0 and the generation current I_g which determine the downward slope of the characteristic for germanium and silicon diodes, respectively, increase with increasing temperature according to Eq. (11-6). As these currents rise, the total reverse current increases, too, and the upward slope of the characteristic shifts to the left and becomes steeper. In germanium diodes, the reverse current almost doubles for each ten degrees of temperature increase. In silicon diodes, the reverse current increases about 2.5 times for each ten degrees of temperature rise.

The effect of temperature variations on the forward current can be attributed to the processes in the junction itself (variations in I_0, I_g and I_r) and changes in the base resistance. The latter results in an increased slope of the characteristic. As a rule, the conductivity of the base material increases with increasing temperature. Consequently, the voltage drop across the base decreases, and the ohmic portion becomes steeper.

The relation between the upward slope of the characteristic and temperature is expressed in terms of the *temperature coefficient of voltage*

$$\varepsilon = \Delta V/\Delta T \qquad (11\text{-}26)$$

which represents the change in forward voltage, ΔV, caused by one degree's change in the temperature.

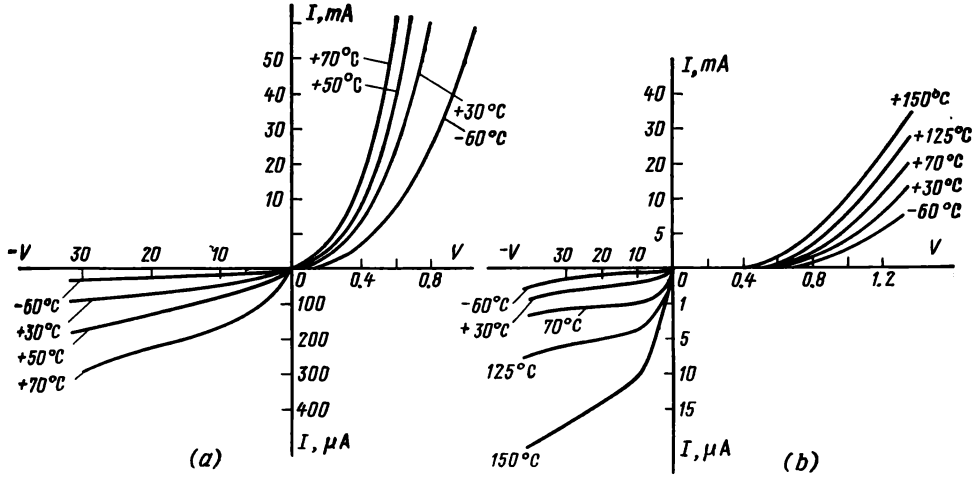


Fig. 11-5. Effect of temperature variations on the volt-ampere characteristics (*a*) germanium diode; (*b*) silicon diode

For silicon and germanium diodes, the temperature coefficient of voltage is approximately equal to —2 mV/°C.

11-4. THE MAIN PARAMETERS OF CRYSTAL DIODES

The volt-ampere characteristic of a crystal diode (as well as that of a vacuum diode) is a nonlinear relation between current and voltage. In the general case, the Q-point of a crystal diode may be positioned by a direct voltage and by an alternating voltage; in the latter case, though, it determines the path of the operating point. So, in addition to the static (d.c.) parameters, the performance of crystal diodes can be expressed in terms of dynamic (a.c.) parameters in much the same manner as that of vacuum diodes.

Dynamic Resistance. This parameter

$$r_{dyn} = \partial V/\partial I \qquad (11\text{-}27)$$

defines variations in the current through a crystal diode when the voltage varies about a certain value, V, associated with the Q-point.

The dynamic resistance is numerically equal to the reciprocal of the slope of the characteristic at the operating point (Fig. 11-6).

For an idealized diode, r_{dyn} can easily be derived by differentiating Eq. (11-22)

$$r_{dyn} = \partial V/\partial I = kT/e\,(I + I_0) \tag{11-28}$$

The dynamic resistance depends on current or applied voltage. When V is below zero, r_{dyn} is high, being from several tens of kilohms to hundreds of megohms.

When the diode is forward-biased, the dynamic resistance decreases with increasing current. In this case, r_{dyn} for real diodes may approximately be found from an equation given in reference [24]

$$r_{dyn} = (\text{approx.})\ 26/I \tag{11-29}$$

where I is in milliamperes and r_{dyn} is in ohms. The above relation is derived from Eq. (11-28) at $I \gg I_0$ and $kT/e = 0.026$ V.

Static (D.C.) Resistance. This is defined as the ratio of the applied direct voltage to the respective direct current

$$r_d = V/I \tag{11-30}$$

Numerically, r_d is equal to the cotangent of the angle that the straight line drawn from the origin of coordinates to the operating point makes with the x-axis (Fig. 11-6). For an idealized diode

$$r_d = V/I_{rev}\,[\exp(eV/kT) - 1] \tag{11-31}$$

or

$$r_d = kT \log_e (I/I_0 + 1)/eI \tag{11-32}$$

For the forward current, the static (d.c.) resistance is designated as $r_{d,\,f}$, for the reverse current, $r_{d,\,r}$.

As a rule, for real diodes, $r_{d,\,f} > r_{dyn}$ and $r_{d,\,r} < r_{dyn}$.

Diode Capacitances. The main parameters also include the barrier and diffusion capacitances. In Sec. 10-6, we have already discussed the capacitances of a P-N junction. Equations (10-66) and (10-68) quite accurately define these capacitances in a crystal diode as functions of voltage and current, respectively

$$C_b = (\text{approx.})\,(\varepsilon\varepsilon_0 s/l)\sqrt{\varphi_c/(\varphi_c + |V|)}$$

$$C_D = (\text{approx.})\,(e\tau_p I/kT)\,[1 - \operatorname{sech}(w_n/L_p)]$$

For a thick-base diode, $\operatorname{sech}(w_n/L_p) = (\text{approx.})\ 0$, so the latter expression reduces to Eq. (10-69)

$$C_D = (\text{approx.})\ eI\tau_p/kT$$

The crystal diode capacitances are very important when diodes are used for pulse work or to convert r.f. or microwave signals (detecting, mixing, etc.). Here, we have to deal with the *total capacitan-*

ce of the diode, C_d. This is the capacitance between the diode leads, as measured at specified bias voltage and frequency. C_d includes not only C_b and C_D, but also the diode case capacitance, C_c.

11-5. CRYSTAL DIODE TRANSIENTS

In some circuits, mainly in computers, crystal diodes are used for pulse work. Here, the current or voltage applied to a diode changes in magnitude and polarity in a very short time, down to 10^{-10} or 10^{-12} s. So, for practical purpose, the voltage or current applied to a diode is a step function related to the internal resistance of the signal source, r_{in}.

Obviously, as the diode is switched from forward to reverse bias, the current through the diode must change in magnitude and direction, too. In a real diode, these changes cannot occur instantaneously because they are associated with the dissipation of the charge constituted by the carriers injected into the base and the recharge of the junction capacitance, both of which introduce a time lag. Therefore, when the voltage applied to a diode is a step function, the current attains its steady-state value only after a certain period of time. Let us discuss these processes, known as *transients*, in more detail.

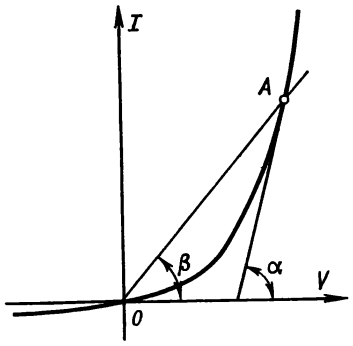


Fig. 11-6. Determining the parameters from the volt-ampere characteristic

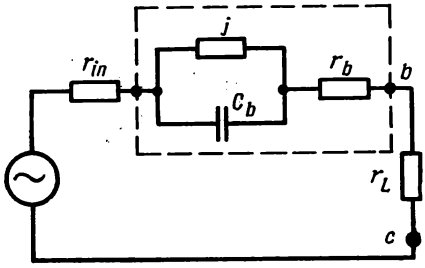


Fig. 11-7. Equivalent circuit for a diode connected to a signal source

Let a signal source be connected to the diode as shown in Fig. 11-7. Here, the diode is represented by an equivalent circuit containing the junction resistance, r_j, the base spreading resistance, r_b, and the barrier capacitance, C_b, providing a by-pass around the junction. Taken together, these elements make up the diode impedance, Z_d, which, in the general case, is a complex quantity: $Z_d = R_d + iX_d$.

If $r_{in} \gg R_d$, the signal source should be regarded as a generator of current whose value depends on r_{in}. When $r_{in} \ll R_d$, the signal source acts as a voltage generator.

The circuit of Fig. 11-7 also contains a low-value instrument load resistance, r_L ($r_L \ll R_d$). If we connect an oscilloscope to the terminals of this resistance at points b and c, we may observe a time plot, or waveform, of the current in the diode, $i_d(t)$. A time plot of the voltage across the diode can be obtained by connecting an oscilloscope to points a and b.

Let us consider the waveforms of the voltage across, and the current in, the diode, $i_d(t)$ and $v_d(t)$, respectively, when a voltage or a current generator is connected to the diode, that is, when the diode is operating at a low and a high injection level.

Figure 11-8 shows the waveforms of the square current pulses and the voltage step input applied to the diode from the current or vol-

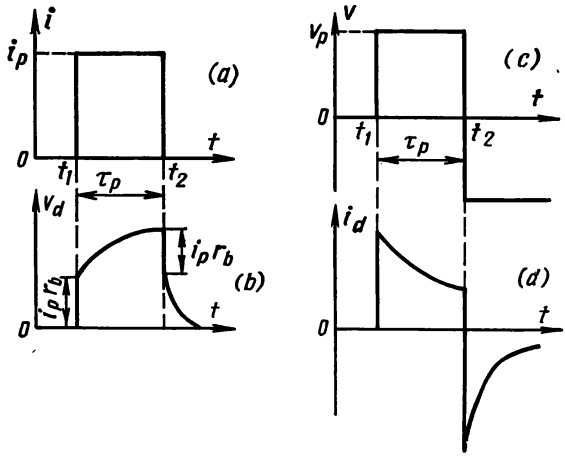


Fig. 11-8. Transients in a diode at a low injection level

(*a*) input current pulse (excitation); (*b*) output voltage (response); (*c*) step-input voltage (excitation); (*d*) output current (response)

tage generator, respectively. Below them are the waveforms of the voltage across, and the current in, the diode when it is assumed to be operating at a low injection level.

Under the circumstances ($\delta \ll 1$), the number of carriers injected into the base is relatively small. So the accumulation and dissipation of carriers in the base and the base pinching are of minor importance. The manner in which the current and voltage attain their steady-state values is mainly determined by the charging and discharging of the barrier capacitance of the *P-N* junction.

When a current pulse is applied to the diode (Fig. 11-8*a*), the current at time t_1 mainly flows through C_b, so v_d is determined by the voltage drop across r_b. At the same time, C_b begins to charge through r_{in} and r_b, and the base voltage is supplemented by the linearly rising voltage across C_b and, as a consequence, the *P-N* junction.

At time t_2 when the current pulse ceases, the voltage immediately decreases by $i_p r_b$, then C_b discharges through r_j, and v_d decreases exponentially.

When the voltage generator is connected to the diode, the transients are likewise attributed to the recharge of C_b. At time t_1, the current i_d is only limited by r_b. As C_b charges, the voltage across the junction rises, and i_d falls exponentially, tending to a steady-state value.

When, at time t_2, the diode is abruptly switched from forward to reverse bias, the reverse current through the diode is rather high; at first, its value depends only on r_b because the voltage across the junction (the capacitance C_b) cannot change instantaneously. Instead, it changes exponentially as C_b recharges, and i_d tends to a steady-state value.

With a high injection level ($\delta \gg 1$), the waveforms of v_d and i_d are somewhat different (Fig. 11-9). Now, the accumulation and dissipation of the carriers injected into the base become predominant.

In addition to the waveforms of current and voltage, Fig. 11-9 also shows the concentration of minority carriers in the base at various moments of time. At time t_1, the voltage v_d is determined by the voltage drop across $r_b + r_j$. Because the time during which the junction capacitance recharges may be neglected, it may be thought that at time t_1, the voltage across the diode reaches v_{d1} decided by r_b and r_j, and holes are injected into the base (curve *1* in Fig. 11-9*c*). Henceforward (time t_2 and t_3), the holes gradually diffuse into the bulk of the base (curves *2* and *3* in Fig. 11-9*c*). As this takes place, the hole concentration gradient at the junction boundary in the base remains constant, because the current through the diode, equal to i_p, remains unchanged.

As a result of the hole storage in the base, its resistance gradually decreases to r'_b during the interval t_1-t_4, and, consequently, the voltage v_d goes down.

When the current is turned off (time t_4), v_d instantly falls by ir'_b. The voltage across the junction cannot rise suddenly because the dissipation of carriers in the base introduces a time lag. Due to the reverse motion across the junction into the emitter and also because of the recombination, the stored hole charge dissipates in the base only during the time t_4-t_6. Now the hole concentration gradient at the junction boundary is zero because $i = 0$.

Similar processes take place when the diode is driven by a voltage generator. After the voltage has been turned on, the current through the diode takes some time to reach a steady-state value because of the carrier injection and storage in the base. At time t_1, the voltage across the diode is constituted by the voltage drop across the P-N junction and the voltage drop across the base resistance which is relatively high because, at this moment, the base contains very few injected carriers (curve *1* in Fig. 11-9*f*). Since v = constant, the vol-

tage across the junction is low and the excess hole concentration at the junction boundary in the base is insignificant. Now i_d is determined by r_d and r_j. Subsequently (at times t_2 and t_3), holes diffuse into

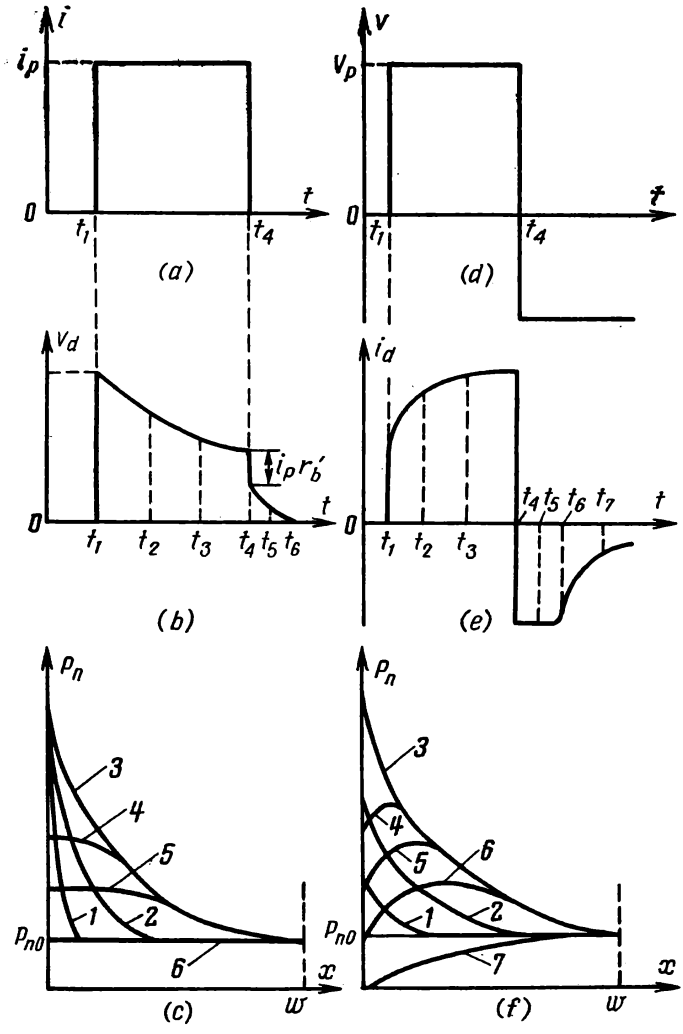


Fig. 11-9. Transients in a diode at a high injection level

(*a*) input current pulse (excitation); (*b*) output voltage (response); (*c*) variations in the carrier concentration in the base; (*d*) step input voltage (excitation); (*e*) output current (response); (*f*) variations in the carrier concentration in the base

the base, its resistance decreases, the voltage across the junction increases, and the excess hole concentration at the junction boundary builds up (curves *2* and *3* in Fig. 11-9*f*). This leads to an increase in the hole concentration gradient and the current i_d.

When the diode is switched from forward to reverse bias, the reverse current reaches a considerable value limited solely by the base resistance which, at this moment, is sufficiently small due to a high concentration of injected carriers. During the time interval from t_4 to t_6, the reverse current remains practically unchanged because the hole concentration gradients at the junction boundary and the base terminal are constant (curves *4* through *6* in Fig. 11-9*f*). After that, the hole concentration, p_n, falls to zero because of the extraction of holes through the junction, and the current i_d gradually decreases to a steady-state value.

11-6. RECTIFIER DIODES

Since crystal diodes are devices with unidirectional conduction, it is logical to use them as rectifiers, that is, devices converting an alternating current into a unidirectional current. Rectifier diodes differ in the source material used (germanium, silicon, etc.), and according to the maximum forward current (low-, medium- and high-power diodes) specified.

The Parameters and Ratings of Rectifier Diodes. The performance of rectifier diodes is expressed in terms of their static (d.c.) parameters discussed in Sec. 11-4 and the ratings which specify their operation in rectifier circuits. The ratings include the average forward current over a cycle, $I_{f,\,av}$; the average reverse current over a cycle, $I_{r,\,av}$; the average rectified current over a cycle (with allowance for the reverse current), $I_{rec,\,av}$; the average forward voltage over a cycle at a specified average forward current, $V_{f,\,av}$; and the average power dissipation, $P_{d,\,av}$, which is the average power dissipated by the diode over a cycle under conditions of forward and reverse currents.

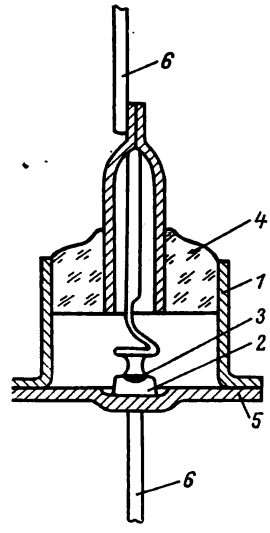


Fig. 11-10. Alloy germanium diode

1—covar case; *2*—germanium crystal; *3*—*P-N* junction; *4*—glass insulator; *5*—support; *6*—leads;

Germanium and Silicon Junction Diodes. As an example, Fig. 11-10 shows a germanium junction diode. Germanium junction diodes are manufactured by alloying a dot of indium into a wafer of *N*-type germanium, and silicon diodes, by alloying a dot of aluminium into a wafer of *N*-type silicon. The lower part of the wafer is soldered to a support or a case. The ohmic contact is provided by a solder containing antimony. Sometimes, high-power silicon diodes are manufactured by diffusing

impurities (boron or phosphorus) into a crystal of *N*- or *P*-type silicon.

As a rule, diodes are enclosed in an air-tight case so that they can operate under conditions of elevated humidity. The maximum operating temperature is +70°C for germanium diodes, and +150°C for silicon diodes. For better heat abstraction, high-power diodes use massive metal cases to which semiconductor wafers are directly soldered. Still more powerful diodes are equipped with external heat sinks or have arrangements for forced air or liquid cooling.

Figure 11-5 shows the volt-ampere characteristics of germanium and silicon diodes and their variations over the operating frequency range. A comparison of these characteristics reveals the typical differences between germanium and silicon diodes. In silicon diodes, the reverse current is much lower, the voltage drop under conditions of forward biasing is generally higher, the maximum forward current density is higher, and the peak reverse voltage is considerably higher than in germanium diodes, being 1500 V instead of 400 V.

They also differ in the effect of temperature on the form of breakdown. In germanium diodes, thermal breakdown is most frequent, and in silicon diodes, avalanche or Zener breakdown.

11-7. CRYSTAL DIODES FOR PULSE WORK

In many cases, crystal diodes are used for pulse work, that is, in pulse-forming and pulse-converting devices and also in switches and logic circuits. These diodes generally have a small junction area (Fig. 11-11). Owing to this, the junction capacitance can be considerably reduced (down to several picofarads) which is especial-

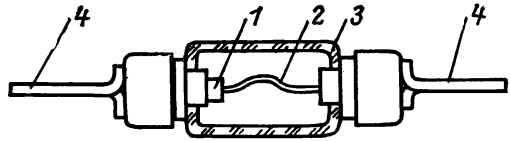


Fig. 11-11. Germanium diode for pulse work

1—germanium crystal; *2*—tungsten catwhisker; *3*—glass case; *4*—leads

ly important for shortening the transient time of diodes. The reduced junction area, however, results in a low power dissipation which ranges between 20 and 30 mW.

The Parameters and Ratings. The main parameters of diodes intended for pulse work are quantities characterizing the transient response to fast variations in the applied voltage or current. These parameters are illustrated in Fig. 11-12. The physical processes underlying the waveforms have been discussed in Sec. 11-5.

The time that passes from the instant when a forward-voltage pulse is initiated until it quiets down to its steady-state value (1.2 $V_{f\text{-}p}$) is called the *turn-on time*, $t_{t\text{-}on}$. The turn-on time depends on the average time of the diffusion of injected carriers in the base, during which the base resistance decreases.

When the applied voltage pulse reverses in polarity, the dissipation of the excess injected carriers in the base due to the diffusion and recombination does not occur instantaneously. This process is

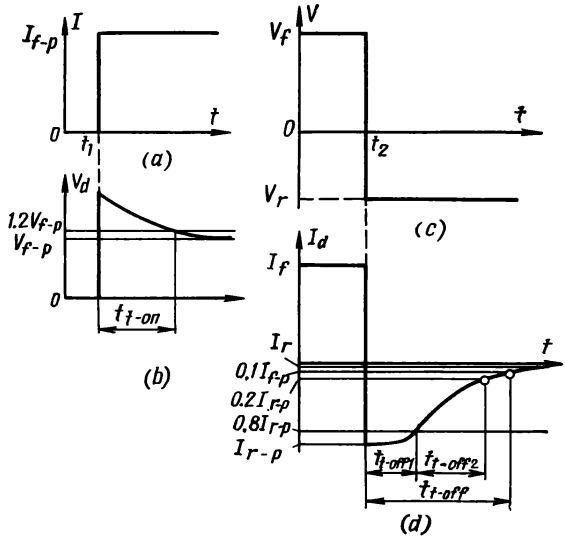


Fig. 11-12. Transients in a pulse diode

(*a*) input current pulse (excitation); (*b*) output voltage (response); (*c*) step input voltage (excitation); (*d*) output current (response)

characterized by what is known as the *turn-off time*, $t_{t\text{-}off}$, which is the time that passes from the instant t_2 at which the diode is switched from forward to reverse bias until the reverse current quiets down to its steady-state value (0.1 $I_{f\text{-}p}$).

Two stages can be differentiated in the turn-off process, namely the stage at which the reverse conduction is high (the interval $t_{t\text{-}off1}$) and the stage at which the reverse current rapidly decreases (the interval $t_{t\text{-}off2}$). These intervals are measured between the instants when the current reaches 0.8 $I_{r\text{-}p}$ and 0.2 $I_{r\text{-}p}$.

The turn-off time is associated with another parameter of importance in pulse work, *the switching charge*, Q_s. This is equal to the total charge flowing through the diode during the transient time, that is, from the instant t_2, when the diode is switched, until the reverse current reaches its steady-state value.

To shorten the turn-off time, the base in some diodes is alloyed by impurities forming traps and contributing to the recombination of

minority carriers. For example, alloying by gold reduces the turn-off time to about 10^{-9} s.

Sometimes, data sheets for diodes intended for pulse work also specify the maximum forward pulse voltage $V_{f\text{-}p,\,max}$, the maximum pulse current, $I_{f\text{-}p,\,max}$, and their ratio known as the *pulse resistance.*

The junction capacitance of diodes specially designed for pulse work should be as low as possible. As a rule, it ranges between a few tenths of a picofarad and several picofarads.

According to $t_{t\text{-}off}$, diodes for pulse work are classed into *millisecond* ($t_{t\text{-}off} > 0.1$ ms), *microsecond* (0.1 ms $> t_{t\text{-}off} > 0.1$ μs) and *nanosecond* ($t_{t\text{-}off} < 0.1$ μs) types.

Snap-Off (Charge-Storage) Diodes. Sometimes, for example, in circuits forming square pulses with sharp leading edges, the ratio $t_{t\text{-}off1}/t_{t\text{-}off2}$ of the diodes should be as small as possible. Here, the diode base is nonuniformly alloyed along its length. In these diodes, called *snap-off* (or *charge-storage*) diodes, the concentration of impurities in the base monotonically increases with increasing distance from the junction. As a consequence, the concentration of majority mobile carriers is nonuniform, too. In the base, electrons diffuse from the N-type semiconductor toward the junction and leave behind positive impurity atoms. As a result, an electric field is developed in the base, whose strength vector is directed toward the junction. Under the action of this field, the holes injected into the base drift back to the junction where they are crowded against the barrier region boundary and form a high-density space charge. When the diode is switched from forward to reverse bias, these holes are pulled in by the junction field during a short time. Because of this, $t_{t\text{-}off2}$ in such diodes is much shorter than in diodes with uniform bases.

Schottky Diodes. These are named after the German physicist Schottky who in the 1930's formulated the principles of the metal-semiconductor junction (*Schottky barrier*).

As a rule, the semiconductor in a Schottky diode is N-type silicon, and the metal electrode is molybdenum, gold, aluminium or some other metal whose work function exceeds that of silicon so that a rectifying junction can be formed. The energy-band diagram of this junction has been discussed in Sec. 10-4 (see Fig. 10-8).

The construction of the Schottky diode is shown in Fig. 11-13. A thin layer (a few micrometres) of high-resistance silicon with an impurity concentration of about 10^{16} cm^{-3} (N-region) is epitaxially grown on a wafer of low-resistance silicon (N^+-region). On the surface of this layer, a layer of metal is deposited by vacuum evaporation. The junction area is usually very small, being from 20 to 30 μm in diameter, and the barrier capacitance does not exceed 1 pF.

A distinction of the physical processes in the Schottky diode is that no minority carriers are injected into the base (silicon). As it

has been shown in Fig. 10-8, the barrier region is formed because the semiconductor layer near the junction is depleted of majority carriers (in our case, electrons). So, when the diode is forward-biased (the positive voltage is applied to the metal), the forward current is constituted by majority carriers (electrons) flowing from the semiconductor across the reduced potential barrier ($\varphi_c - V$) into the metal. As is seen, no storage and diffusion of minority carriers take place in the diode base (N-type silicon). The transient time is

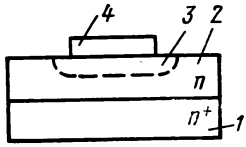


Fig. 11-13. Schottky diode
1 — low-resistance N+-type silicon; *2*—epitaxial layer of high-resistance N-type silicon; *3*—barrier region; *4*—metal terminal

mainly affected by the recharging of the barrier capacitance. As already noted, C_b is very low, being not higher than 1 pF; the resistances of the metal and N^+-type silicon are also very small. Because of this, the time necessary for C_b to recharge and, as a consequence, the transient time are very short, too, being a few tenths of a nanosecond. This makes it possible to use Schottky diodes in nanosecond switches, and also at frequencies up to 10 or 15 GHz.

The volt-ampere characteristic of the Schottky diode is almost ideally described by the exponential relation given by Eq. (10-52) for an idealized diode. Owing to this, Schottky diodes are successfully used in logarithmic circuits.

High-power Schottky diodes intended for operation in rectifier circuits may handle forward currents up to several tens of amperes at a forward voltage of 0.5 to 1 V. The peak reverse voltage in such diodes reaches 200 to 500 V.

The reverse current in the Schottky diode is low; in switching diodes, I_r is as low as a few picoamperes. The reverse current depends on the equilibrium concentration of electrons near the junction, n_0, their mean thermal velocity, v_m, and the junction area, s.

11-8. MIXER AND DETECTOR DIODES

Mixer Diodes. These are intended to convert r.f. signals into intermediate-frequency signals in superheterodyne receivers. So, they carry out the same functions as frequency-mixer tubes.

Mixer diodes (Fig. 11-14) are of the point-contact type so that they can be built directly into coaxial lines or waveguides. Owing to a low interelectrode capacitance (which is below 1 pF), these dio-

des are used to advantage in the frequency ranges up to tens of gigahertz.

The Parameters and Ratings of Mixer Diodes. Of practical importance are the parameters describing the diode as a mixer stage in a receiver. One of them is the *conversion loss*

$$L_{con} = 10 \log_{10} (P_{rf}/P_{if}) \qquad (11\text{-}33)$$

where P_{rf} is the power of the r.f. signal at the input of the mixer stage and P_{if} is the power of the intermediate-frequency signal.

The efficiency of frequency conversion by a diode mixer depends on the nonlinearity of its volt-ampere characteristic. Therefore, the conversion loss depends on which portion of the volt-ampere characteristic is used and, as a consequence, on the current flowing through the diode. L_{con} ranges from diode to diode between 5 and 10 dB.

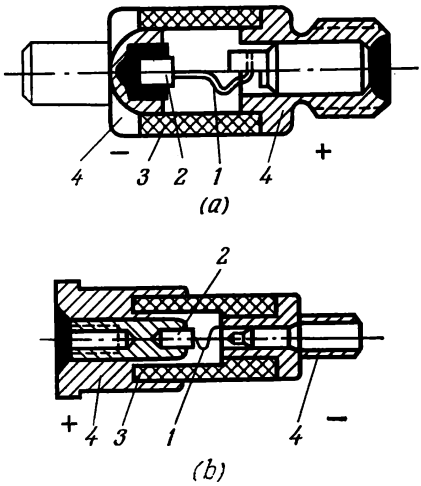


Fig. 11-14. Radio-frequency (r.f.) diodes

(*a*) silicon; (*b*) germanium; *1*—metal catwhisker; *2*—semiconductor crystal; *3*—ceramic case; *4*—metal flanges serving as leads

Another important parameter of mixer diodes is the *noise temperature*, t, defined as the ratio of the diode noise power to the thermal noise power in an equivalent resistance at room temperature

$$t = P_n/kT\,\Delta f \qquad (11\text{-}34)$$

The noise temperature is determined at a power of 1 mW supplied by a local oscillator. As a rule, t is approximately equal to 2 or 3.

Plots of L_{con} and t as functions of the diode current are shown in Fig. 11-15. Since the sensitivity of a receiver is affected not only by noise but also by the conversion loss, the operating rectified current is chosen to minimize the conversion loss and noise as much as possible. For example, with the local oscillator supplying 0.5 or 1 mW of power, the optimum rectified current for silicon diodes is 0.4 mA.

The ratings specified by data sheets for mixer diodes include the *peak pulse power*, $P_{p\text{-}p}$, that is, the maximum safe power which can be applied to the mixer input without damaging it. For silicon diodes, $P_{p\text{-}p}$ ranges between 30 and 80 mW, and for germanium diodes, it is about 150 mW.

Another important rating is the *input impedance*, Z_{in}, because failure to match the mixer input impedance to that of the input circuit may result in the reflection of a considerable part of the incoming signal power. The input impedance of the diode depends

not only on the physical characteristics of the *P-N* junction (such as the dynamic resistance, r_{dyn}, junction capacitance, C_j, and base resistance, r_b) but also the distributed reactances: the case capacitance and the catwhisker inductance. So for matching purposes, when designing a diode, one seeks to reduce Z_{in} to its resistive component equal to the standard characteristic impedance of the transmission line (50 or 75 ohms).

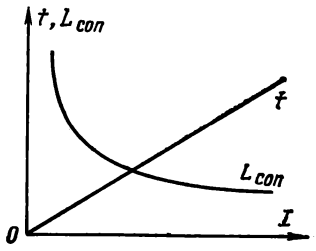


Fig. 11-15. Conversion loss and noise temperature as functions of current in a mixer diode

Detector Diodes. These are intended to detect radio signals in radio receivers and various instruments.

The parameters of detector diodes include *current* and *voltage sensitivities.*

Current sensitivity is defined as the ratio of the incremental change in the rectified current to the power of the applied r.f. signal that causes the change:

$$\beta_I = \Delta I_{rec}/P \qquad (11\text{-}35)$$

Voltage sensitivity is defined as the ratio of the small change in the output voltage produced by the power of the applied r.f. signal:

$$\beta_V = \Delta V_{out}/P \qquad (11\text{-}36)$$

Generally, the upward slope of the volt-ampere characteristic of detector diodes is well described by the square current-voltage relation. The breakdown voltage in these diodes is rather low because the base is highly-alloyed in order to reduce r_b.

For detecting large signals, detector diodes should have a higher breakdown voltage, which runs somewhat counter to the requirement for a reduced r_b. In such a case, use is mostly made of Schottky diodes (see Sec. 11-7).

11-9. ZENER DIODES

Zener diodes are intended to operate as voltage stabilizers, voltage reference, and so on. They are predominantly made of *N*-type silicon. This is because silicon diodes have a low reverse current, experience an avalanche or a Zener breakdown at insignificant variations in the reverse voltage, and have a high maximum safe temperature of the *P-N* junction.

The volt-ampere characteristic of the Zener diode (Fig. 11-16*a*) is the breakdown region within the reverse slope of the volt-ampere characteristic of the *P-N* junction. As has been shown in Sec. 10-5, the avalanche or Zener breakdown voltage depends on the resistivity of the semiconductors making up the *P-N* junction. Using slabs of

N-type silicon with different impurity concentrations, we can manufacture Zener diodes with different V_{bd} and, as a consequence, with different values of voltage stabilization. For highly-alloyed N-type silicon, the Zener breakdown is predominant at $V_{bd} \leqslant 6$ V; at V_{bd} = (approx.) 5 to 7 V, the Zener breakdown is accompanied by the avalanche breakdown, which becomes predominant at $V_{bd} > > 7$ V.

Like the gas-filled stabilizer tube (Sec. 8-4), the Zener diode operates on the principle that the voltage across the diode terminals (in the breakdown region) varies insignificantly with considerable variations in current. The Zener diode is connected in a circuit as shown in Fig. 11-16*b*. When, say, the supply voltage, E, increases, it would appear that the current in the common circuit, I, the current in the load resistor, and the voltage drop across the resistor, R_L, should increase. Actually, however, the excess current I in the common circuit is absorbed by the Zener diode (I_Z increases), whereas the voltage across the diode and, as a consequence, load resistor R_L remains the same.

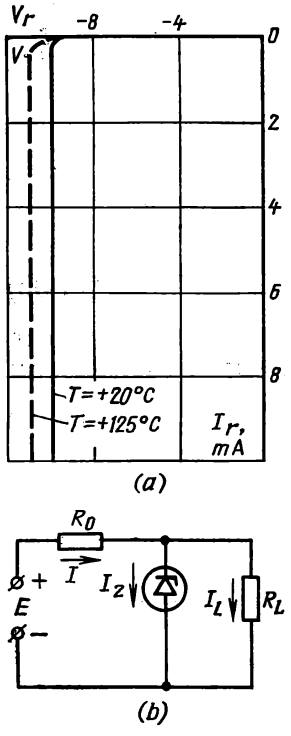


Fig. 11-16. (*a*) Characteristic of a Zener diode and (*b*) its connection in a circuit

The Parameters and Ratings of Zener Diodes. One of the main ratings specified by data sheets for Zener diodes is the stabilization (maintaining) voltage, V_s. As its name implies, this is the voltage stabilized by the diode at a specified stabilization current. According to V_s, Zener diodes are divided into low-voltage and high-voltage types. In the Soviet Union, Zener diodes handling voltages from 3 to 400 V are available.

The ratings also include the *maximum* and *minimum stabilization currents*, I_{max} and I_{min}. These represent the limits of the region in the volt-ampere characteristic of a Zener diode, which can be used for voltage stabilization with a desired reliability.

The parameters of Zener diodes are the *dynamic resistance*

$$r_Z = \Delta V_s / \Delta I \tag{11-37}$$

and the *static* (*d.c.*) *resistance*

$$R_{st} = V_s / I \tag{11-38}$$

Here, V_s and I are the voltage and the current at a specified quiescent point, and ΔV_s and ΔI are the incremental changes in these quantities.

Referring to Fig. 11-16, the stabilization voltage varies with temperature, which fact is expressed in terms of the *temperature coefficient of voltage*

$$\alpha_s = (1/V_s)\,(\Delta V_s/\Delta T) \qquad (11\text{-}39)$$

where ΔV_s is the change in the stabilization voltage caused by a change ΔT in temperature.

11-10. VARICAPS

Also known as *varactors*, these devices are *P-N* junction crystal diodes whose junction capacitance varies with the applied reverse voltage. Varicaps belong to the class of parametric devices. They are used in frequency multipliers and microwave parametric amplifiers. However, these applications of varicaps will not be discussed in this text, and the reader should turn to a course on microwave and quantum devices.

The performance of a varicap is expressed by its *volt-farad characteristic*, $C_v = f(V_r)$, where C_v is its total capacitance, that is, the capacitance measured between its leads. In addition to the barrier capacitance, C_b, the total capacitance, C_v, includes the capacitance of the varicap case, C_c. Since $C_b \gg C_c$, the volt-farad characteristics of varicaps are similar to those of the *P-N* junction shown in Fig. 10-13. Referring to these curves, the relation $C_j = f(V_r)$ is seen to depend on the type of junction, being most pronounced for a junction whose impurity concentration is described by a complex function.

The Parameters and Ratings of Varicaps. The relation $C_v = f(V_r)$ is characterized by the *capacitance ratio*

$$K_C = C_{v1}/C_{v2} \qquad (11\text{-}40)$$

where C_{v1} and C_{v2} are the total capacitances of the varicap at specified values of the reverse voltage V_{r1} and V_{r2}.

Sometimes, the nonlinearity of the volt-farad characteristic is expressed in terms of the *nonlinearity factor*

$$K_n = \Delta C_v/C_v\,\Delta V_r \qquad (11\text{-}41)$$

These two parameters are interrelated because an increase in the nonlinearity of the volt-ampere characteristic may cause the capacitance to change from C_{v1} to C_{v2} with a smaller variation in the reverse voltage. For example, in varicaps utilizing an alloy *P-N* junction K_C reaches 10 when the reverse voltage changes from zero to a few

tens of volts. In varicaps having a sharply rising volt-farad characteristic, a change of 10 V in the reverse voltage results in K_C = = (approx.) 100.

The quality of a varicap is expressed in terms of its *quality factor*, Q_v (or, simply, *Q-factor*), which is equal to the ratio of the varicap

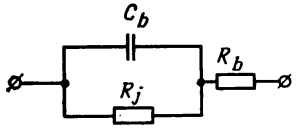


Fig. 11-17. Equivalent circuit for a varicap

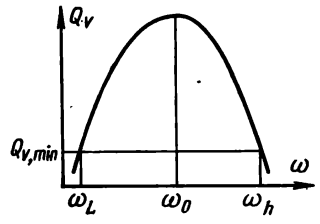


Fig. 11-18. *Q*-factor of a varicap as function of frequency

reactance at a specified frequency to the resistance loss at a specified capacitance.

Referring to the equivalent circuit for a varicap shown in Fig. 11-17, we may write for the *Q*-factor

$$Q_v = \omega C_b/[1/r_j + r_b(1/r_j^2 + \omega^2 C_b^2)] \qquad (11\text{-}42)$$

As is seen, the *Q*-factor depends on frequency. This relation is plotted in Fig. 11-18.

Differentiating Eq. (11-42) with respect to ω and equating the derivative to zero, we get the frequency ω_0 corresponding to the maximum *Q*-factor

$$\omega_0 = (1/C_b)\sqrt{(1/r_b r_j)(1 + r_b/r_j)} \qquad (11\text{-}43)$$

Using this expression, we get the maximum *Q*-factor

$$Q_{v,\max} = (1/2)\sqrt{1/(r_b/r_j)(1 + r_b/r_j)} \qquad (11\text{-}44)$$

In real varicaps, the ratio r_b/r_j = (approx.) 10^{-7}. In view of this, Eqs. (11-43) and (11-44) can be reduced to

$$\omega_0 = (\text{approx.})(1/C_b)\sqrt{1/r_b r_j} \qquad (11\text{-}45)$$

$$Q_{v,\max} = (\text{approx.})(1/2)\sqrt{r_j/r_b} \qquad (11\text{-}46)$$

In real devices, $Q_{v,\max}$ reaches several thousand.

At audio frequencies, r_b may be neglected because, owing to the high reactance of the barrier capacitance, it is much smaller than $1/\omega C_b$ shunted by r_j.

Then,

$$Q_{v,af} = (\text{approx.})\,\omega C_b r_j \qquad (11\text{-}47)$$

At radio frequencies, $1/\omega C_b$ decreases and the shunt resistance r_j may be ignored.

As a result,

$$Q_{v,\,rf} = (\text{approx.})\ 1/\omega C_b r_b \qquad (11\text{-}48)$$

As follows from Eq. (11-47), varicaps designed for audio frequencies should have very high C_b and r_j. To meet this requirement, use is made of materials having a wide forbidden band (a reduced reverse current per unit junction area). At $V = 0$, C_b is several tenths of a microfarad.

In accordance with Eq. (11-48), C_b and r_b for varicaps used at radio frequencies should be minimal. The base resistance, r_b, may be reduced by raising the impurity concentration in the base. This, however, brings about a decrease in the breakdown voltage, which is undesirable. The breakdown voltage may be raised by using materials with a high carrier mobility.

The operating frequency range is stated in terms of the upper (ω_{up}) and lower (ω_l) frequencies measured at a minimum Q-factor, $Q_{v,\min}$. The minimum Q-factor is usually taken to be unity. However, varicaps with $Q_{v,\min} = 1$ are unsuitable for parametric circuits. So, in such cases, $Q_{v,\min}$ is required to be more than unity, for example, $Q_{v,\min} = 10$. The upper frequency, ω_{up}, corresponding to $Q_{v,\min} = 1$ is often called the *critical frequency*

$$\omega_{cr} = 1/r_b C_b \qquad (11\text{-}49)$$

The varicap ratings considerably depend on temperature, although C_b is insignificantly affected by temperature variations. As the temperature rises, r_j sharply falls because of an increase in the reverse current. Owing to this, $Q_{v,af}$ goes down significantly. Varicaps operate satisfactorily only at relatively low temperatures. The operating temperature of varicaps made of gallium arsenide is $T_{op} \leqslant \leqslant 150°C$, and that of germanium varicaps ranges between 50 and 60°C.

The relation between the varicap parameters and temperature is expressed in terms of the *temperature coefficient of capacitance*

$$\alpha_{C_v} = \Delta C_v / C_v \Delta T \qquad (11\text{-}50)$$

and the *temperature coefficient of the Q-factor*

$$\alpha_{Q_v} = \Delta Q_v / Q_v \Delta T \qquad (11\text{-}51)$$

Here, ΔT is the change in ambient temperature.

Chapter Twelve

BIPOLAR TRANSISTORS

12-1. DESIGN AND PRINCIPLE OF OPERATION

Definition. A *transistor* (transfer resistor) is an active semiconductor device with two or more *P-N* junctions and three or more electrodes. It can be used for power amplification.

Transistors are many and diverse. Modern transistors fall into two main classes: *bipolar* transistors and *field-effect* transistors, differing in the principle of operation, main characteristics and parameters.

A bipolar transistor is a device which uses both negative and positive carriers. Because this chapter will only discuss bipolar transistors, they will be referred to as simply transistors. Field-effect transistors will be discussed in Chapter 13.

The construction of various transistors is schematically shown in Fig. 12-1. The base of an alloy transistor (Fig. 12-1*a*) is a wafer of *N*-type germanium sandwiched between two pellets of indium. Indium atoms diffuse into the germanium and form two regions with hole (*P*-type) conduction, known as the emitter and collector regions (or, simply, emitter and collector). The base, emitter and collector are brought out with the help of leads soldered to them. Thus, in the simplest case, the transistor is a three-layer structure in which the outer regions are made up of semiconductors whose conduction differs from that of the middle layer. These regions are separated by *P-N* junctions.

Figure 12-1*b* shows a *planar* transistor in which the electrodes and their leads come to the same plane—the surface of the wafers. The wafer of *N*-type silicon serving as the collector contains a base region (*P*-type silicon) and an emitter region (N^+-type silicon) with an increased donor impurity concentration both manufactured by local diffusion (that is, by introducing alloying material atoms into the semiconductor crystal through a part of its surface).

The interface between the emitter and base regions and that between the base and collector regions make up *P-N* junctions. The

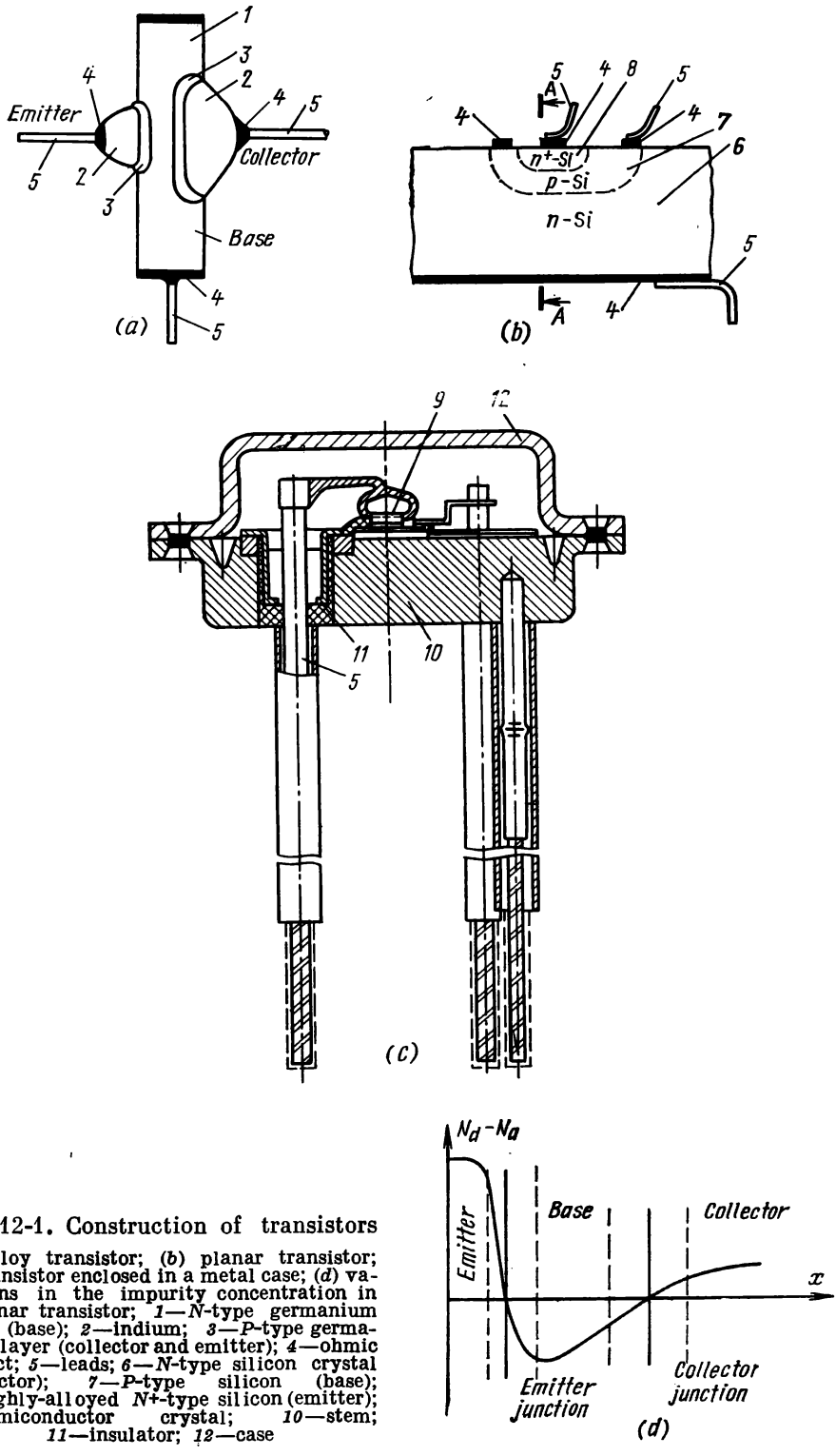


Fig. 12-1. Construction of transistors

(*a*) alloy transistor; (*b*) planar transistor; (*c*) transistor enclosed in a metal case; (*d*) variations in the impurity concentration in a planar transistor; *1*—*N*-type germanium wafer (base); *2*—indium; *3*—*P*-type germanium layer (collector and emitter); *4*—ohmic contact; *5*—leads; *6*—*N*-type silicon crystal (collector); *7*—*P*-type silicon (base); *8*—highly-alloyed *N*+-type silicon (emitter); *9*—semiconductor crystal; *10*—stem; *11*—insulator; *12*—case

lower surface of the transistor is given a coat of metal film to which the collector lead is welded. Similar coats of metal film are applied and the respective leads are welded to the emitter and base regions.

The silicon crystal is carried by a support and the whole transistor structure is enclosed in an air-tight metal, plastic or glass case (Fig. 12-1*c*); the electrode leads are brought out through insulators in the case bottom.

In real transistors, the emitter, base and collector are alloyed to a different degree. As a rule, the impurity concentration in the emitter is several orders of magnitude higher than in the base. In planar transistors, the base and collector are alloyed to about the same extent. Figure 12-1*d* shows typical variations in the impurity concentration in a planar transistor at section *A-A*.

In alloy transistors, the impurity concentrations in the collector and emitter are about the same.

Classification and Diagram Symbols. Transistors can be classified in several ways. To begin with, they are divided according to the material into germanium and silicon types. Also, transistors may be classed according to the sequence in which the *N*- and *P*-regions are arranged. For example, the alloy transistor shown in Fig. 12-1*a* is a *P-N-P* transistor, and that of Fig. 12-1*b* is known as an *N-P-N* transistor.

Transistors are also classified according to the method of manufacture (alloy, micro-alloy, mesa, diffused, planar, etc.), their power rating (low-, medium- and high-power transistors), operating frequency range (audio-, medium- and radio-frequency transistors), the predominant mechanism of current in the base (drift and diffusion types), and so on.

In the Soviet Union, the diagram symbols for transistors are set up by USSR State Standard GOST 2.728-68, GOST 2.730-68 and GOST 10862-72.

Configurations of Transistor Circuits. As in other electron devices, the transistor electrodes are supplied with direct voltages from external sources. If we take the voltage at one of the three electrodes to be zero, the voltages at the other electrodes can conveniently be reckoned with respect to this datum level. Accordingly, external supply sources can be connected to a transistor in one of three ways (Fig. 12-2).

The arrangement of Fig. 12-2*a* is known as a *common-* or *grounded-base* circuit. In this case, the voltages at the emitter and collector are determined relative to the base voltage which is taken to be zero. In the *common-* or *grounded-emitter* and *common-* or *grounded-collector* configurations, the voltages are determined with respect to the emitter and collector voltages, respectively.

In the overwhelming majority of cases, the emitter or base circuits act as input, because they accept the signals to be converted,

and the collector circuit usually containing a load resistor serves as output. In the common-collector configuration, the emitter circuit is used as output.

Operating Regions. In any configuration, the supply sources may be connected to the common electrode with either the positive or negative terminal. As a result, each of the two *P-N* junctions can be forward- or reverse-biased. Accordingly, transistors can operate in four main regions.

One is the *active region*. In this case, the emitter junction is forward-biased and the collector junction is reverse-biased. Operation

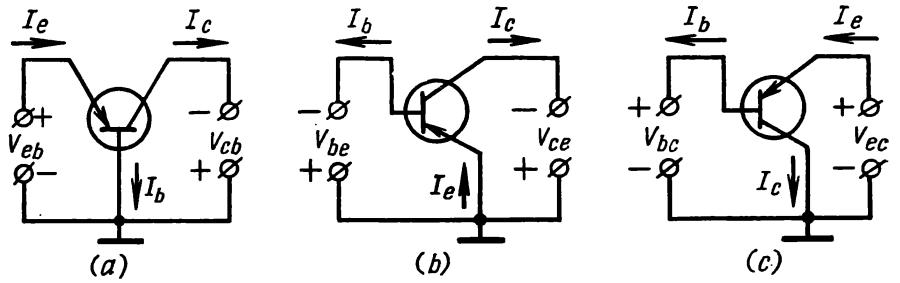


Fig. 12-2. Three configurations of the *P-N-P* transistor

(*a*) common-base connection; (*b*) common-emitter connection; (*c*) common-collector connection

in this region is shown in Fig. 12-2. Another is the *saturation region*. Now both the emitter and collector junctions are forward-biased. The third is the *cut-off region*. Here, both the emitter and collector junctions are reverse-biased. The fourth is the *inversion region*, with the collector junction forward-biased and the emitter junction reverse-biased.

Principle of Operation. Let us discuss in outline the principle of transistor operation using its simplified model. For this purpose, we shall take an alloy *P-N-P* transistor connected in a common-base circuit. We shall assume that the planes of the *P-N* boundaries at the junctions are parallel, the impurity concentrations in the emitter and collector regions are approximately equal ($N_{a,e} \approx \approx N_{a,c}$), and the impurity concentration in the base is much smaller than elsewhere ($N_{d,b} \ll N_{a,e}$).

An energy diagram of an alloy *P-N-P* transistor in the absence of voltages at its electrodes is shown in Fig. 12-3*a*. The entire system is in equilibrium and has a single Fermi level. At room temperature, almost all of the impurity atoms are ionized. Because of this, the Fermi level beyond the boundaries of the barrier regions is located several kT above the local acceptor levels in the *P*-type semiconductors, and several kT below the local donor level in the *N*-type semiconductor.

Since the emitter and collector are highly alloyed, most of the barrier region in the emitter and collector junctions is in the base region having a higher resistivity.

A different energy diagram applies when the transistor is connected in a common-base circuit. In the active region, the emitter is supplied with a positive voltage V_{eb}; the energy bands of the emitter

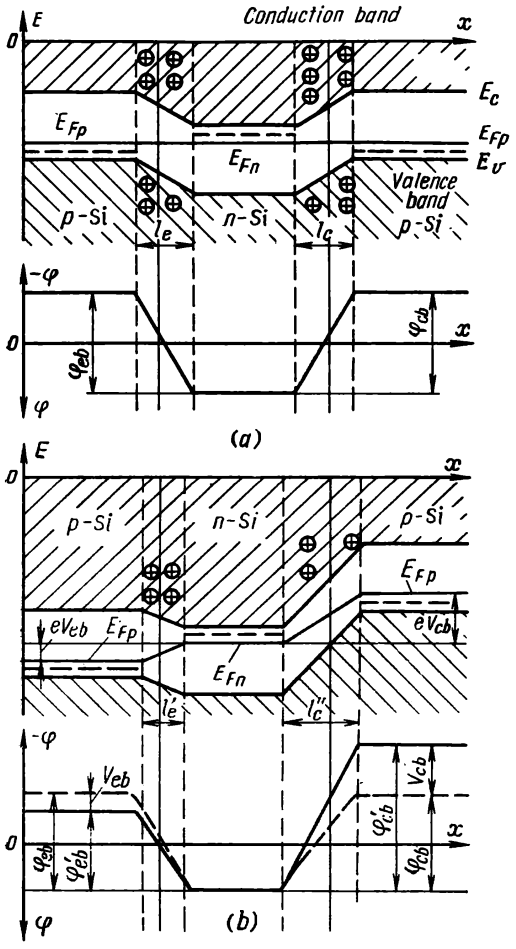


Fig. 12-3. Energy diagrams and potential distribution in the *P-N-P* transistor

(*a*) without external voltages; (*b*) with external operating voltages applied

shift down towards the positive potential region, and the potential barrier at the emitter junction decreases to $\varphi'_{eb} = \varphi_{eb} - V_{eb}$. As a result, the barrier region becomes narrower.

In the active region, the collector-base voltage, V_{cb}, is applied in negative polarity; the energy bands shift up towards the negative potential region, the potential barrier at the collector junction rises to $\varphi'_{cb} = \varphi_{cb} + |V_{cb}|$, and the barrier region widens.

The reduced potential barrier at the emitter junction allows majority carriers to move from the emitter into the base. Because the concentration of holes in the emitter region is greater than that of electrons in the base ($p_{pe} \gg n_{nb}$), the injection ratio, γ, is rather high.

Owing to the injection of holes from the emitter into the base, the hole concentration in the base rises. The positive space charge

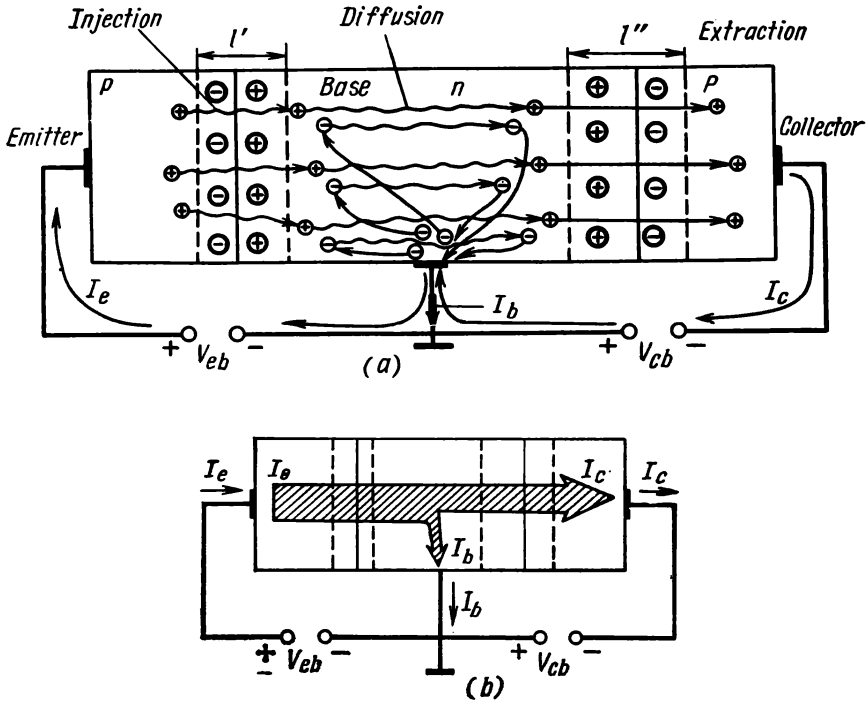


Fig. 12-4. (*a*) Movement of charge carriers and (*b*) currents in a transistor

constituted by holes near the junction in the base is neutralized almost instantly, during the dielectric relaxation time, by the charge due to the electrons supplied to the base by the source of V_{eb}. This completes the circuit for the emitter-base current (Fig. 12-4*a*). The electrons appearing in the base move to and form at the emitter junction a negative space charge which almost entirely neutralizes the charge established by holes. As a result, a region with an increased concentration of electrons and holes is formed near the emitter junction. The difference in concentration gives rise to the diffusion of holes and electrons toward the collector. In transistors, the base width is chosen so that the lifetime of minority carriers (holes) is considerably longer that the time of their motion in the base. Because of this, an overwhelming majority of holes injected from

the emitter (practically about 99% or more) have no time to recombine with electrons in the base. Near the collector junction, the holes find themselves in its accelerating field and are attracted into the collector region—the extraction of holes from the base into the collector takes place. At the same time, electrons whose number is equal to the number of holes which have left for the collector move to the base terminal. As a result, the circuit for the collector-base current is completed.

Thus, the emitter current, I_e (one flowing across the emitter junction) is a control current, and the collector current, I_c (one in the collector circuit) which depends on I_e, is a controlled current. The base current, I_b, is the difference between the control and controlled currents (the hole recombination current in the base). In the base lead, the majority carriers of the base region (electrons) move in opposite directions as they balance out the streams of holes moving through the emitter and collector junctions.

In other words, in the active region, the current flows from the emitter through the base to the collector. Due to recombination in the base, an insignificant part of this current flows into the base circuit (Fig. 12-4*b*).

Generally speaking, the current through a transistor can be controlled by varying the voltage across either of the two *P-N* junctions. In the active region, however, the emitter current and, as a consequence, the collector current do not equally depend on V_{eb} and V_{cb}.

The emitter junction is forward-biased, so the current through this junction and the collector current markedly depend on V_{eb}, increasing exponentially with increasing voltage. Thus, the current in a transistor can easily be controlled over a wide range by varying its emitter voltage.

The relation between the current in a transistor and the reverse collector voltage is somewhat different. Even at $V_{cb} = 0$, the holes which have passed through the base and approached the collector junction will be attracted by the contact field $\mathscr{E}_c$ into the collector region. When reverse collector voltage is applied to the junction, the junction field rises to $\mathscr{E}_\sigma = \mathscr{E}_c + \mathscr{E}_{ex}$, where $\mathscr{E}_{ex}$ is the field developed by V_{cb}. In this case, however, the collector current remains practically unchanged because all of the holes that arrive at the collector junction and whose number only depends on the number of holes injected from the emitter into the base and their recombination in the base will be pulled into the collector region, irrespective of the accelerating field.

Thus, transistors meet the requirements placed upon the electron devices intended to convert electric signals. The current in a transistor can easily be controlled by the incoming signal, and it is insignificantly affected by the output voltage.

As already noted in Chapter 4, the type of electron tube that meets the requirements best is the pentode in which the plate current can easily be controlled by varying the input voltage (that is, one between the control grid and cathode), and the output (plate) voltage has an insignificant effect on the plate current.

From this point of view, the difference between the transistor and the pentode is in the input characteristics. When the pentode control grid is negative and draws no current ($I_{g1} = 0$), the input resistance of the pentode, $R_{in} = dV_{g1}/dI_{g1}$, is infinity. In the transistor, the input resistance, $R_{in} = dV_{eb}/dI_e$, is rather small because, when the emitter junction is forward-biased, the input (emitter) current is markedly affected by the input voltage, V_{eb}. Consequently, in the pentode, the output (plate) current depends on the input voltage applied between the control grid and the cathode. In the transistor, the input voltage, V_{eb}, controls the input current, I_e, which, in turn, affects the output current, I_c. Because of this, the transistor is sometimes called a *current-controlled* device in contrast to vacuum tubes which are *voltage-controlled* devices.

Owing to their properties, transistors may be used in amplifier circuits and, in the case of positive feedback, in oscillators. The operation of the loaded transistor as an amplifier will be discussed in Sec. 12-6.

12-2. PHYSICAL PROCESSES IN THE TRANSISTOR AND ITS PARAMETERS

Most physical processes taking place in transistors and their parameters have been discussed in Chapter 11 as applied to crystal diodes. These include the injection of minority carriers into the base, the ratio and level of injection; some physical processes in the base (such as the base resistance modulation and the base field); the components of the forward and reverse currents through the junction; and the junction capacitances.

In contrast to the crystal diode, however, the transistor has two interconnected *P-N* junctions, so, in addition to the above parameters, the processes occurring in the transistor are characterized by some other quantities. These parameters describe such process as the partition of the emitter current into the base current and the collector current (the relation between these currents), the effect of the voltage across one of the junctions on the processes in the transistor base and the current through the other junction, and so on.

Transistor Currents. As already noted, in the active region, the holes injected from the emitter move in the base and are attracted by the field of the collector junction to constitute the collector current, I_c. Owing to recombination in the base and also for some other reasons, I_c is smaller than I_e. On the basis of Kirchhoff's

current law, we may write for the currents flowing in the various transistor circuits (or leads) thus

$$I_e = I_c + I_b \tag{12-1}$$

In the active region, the emitter junction is forward-biased and it passes I_e which consists of I_{ep} and I_{en} due to the injection of holes from the emitter into the base and electrons from the base into the emitter, and also the recombination current, I_{er} (the leakage current, I_{el}, may be neglected)

$$I_e = I_{ep} + I_{en} + I_{er} \tag{12-2}$$

As a rule, the hole concentration in the emitter is much higher than the electron concentration in the base, $p_{pe} \gg n_{nb}$, so $I_{ep} \gg \gg I_{en}$ (the injection ratio γ is approximately equal to 1). The recombination current, I_{er}, forms a considerable part of the total current only at low forward voltages (for silicon transistors, $V < 0.2$ V). At higher voltages, I_{ep} is much higher than I_{er}. So, we may write for the emitter current

$$I_e = (\text{approx.})\, I_{ep} \tag{12-3}$$

In the active region, the collector current flows through the reverse-biased collector junction. So, in addition to the reverse current which, in the general case, is equal to $I_{rev} = I_0 + I_g + I_l$ according to Eq. (11-11), the collector junction passes the current due to the extraction of holes from the base into the collector

$$I_{cp} = I_{ep} - I_{br} \tag{12-4}$$

which is equal to the hole component of the emitter current minus the current due to the recombination of holes in the base.

When the reverse collector voltage, V_{cb}, is sufficiently high, the carriers may be multiplied by impact or field generation.

Thus, if we ignore I_l, the collector current is

$$I_c = (\text{approx.})\, M\, (I_{cp} + I_{cg} + I_{c0}) \tag{12-5}$$

where M is the multiplication factor for carriers in the barrier region, defined by Eq. (10-59).

As already noted in Sec. 11-2, for silicon devices, $I_{cg} \gg I_{c0}$, so the last term in the brackets in Eq. (12-5) may be neglected.

For normal operating conditions, when V_{cb} is much lower than the breakdown voltage, M is approximately equal to unity and the collector current may be defined as

$$I_c = (\text{approx.})\, I_{cp} + I_{cg} \tag{12-6}$$

The base current may be defined as the difference between the emitter and collector currents.

Reverse (Leakage or Cut-off) Currents. The reverse (also known as leakage or cuttoff) collector (or emitter) current is a current in

the collector (or emitter) junction at a specified reverse voltage across the same junction, with the circuit of the other junction open or, which is the same, for zero emitter (collector) current: $I_e = 0$ (or $I_c = 0$).

Since, under these conditions, $I_{ep} = 0$ and, as a consequence, $I_{cp} = 0$, the reverse (leakage or cutoff) collector current

$$I_{cbo} = I_{cg} + I_{c0} \tag{12-7}$$

due to the generation of carriers in the collector and base regions and in the collector junction is that part of the collector current which is not controlled by the processes at the emitter junction*. The reverse collector current, I_{cbo}, is very important when the transistor operates in the active region, that is, when the collector junction is reverse-biased.

Similarly, the reverse (leakage or cut-off) emitter current, I_{ebo}, is that component of the emitter current whose value depends on the generation of carriers in the emitter and base regions and at the emitter junction. This current is important when the transistor operates in the inverse region, that is, when the emitter is reverse-biased.

In addition to I_{cbo} and I_{ebo} measured with the emitter or collector open-circuited, respectively, it is customary to distinguish also I_{cbs} and I_{ebs}.

The former, I_{cbs}, is the reverse collector current with the emitter short-circuited. The latter, I_{ebs}, is the reverse emitter current with the collector short-circuited.

$I_{cbs} > I_{cbo}$ and $I_{ebs} > I_{ebo}$ because the distribution of minority carriers in the base when the emitter or collector are open-circuited differs from that when the emitter or collector are short-circuited. This will be discussed in more detail in a section devoted to the processes in the base.

Current Gains. Taking into consideration the reverse collector current, the total collector current, I_c in the active region should be treated as the sum of two components, namely I_{cbo} and a part of the emitter current due to the flow of the carriers injected into the base and reaching the collector junction (minus those which have recombined in the base).

So,

$$I_c = \alpha I_e + I_{cbo} \tag{12-8}$$

---

* The subscripts of the currents have the following meaning. The first letter indicates the electrode which passes the current in question; the second letter designates the common electrode; the third letter describes the state of the circuit between the two other electrodes: "o" stands for an open circuit, "s" for a short circuit, "R" for the specified circuit resistance, and "X" for the specified voltage between the electrodes.

The quantity

$$\alpha = (I_c - I_{cbo})/I_e \tag{12-9}$$

is called the *common-base d.c. current gain* (also known as the *alpha current gain*). As a rule, $\alpha < 1$. In the inverse region when the collector junction is forward-biased and the emitter junction is reverse-biased, the emitter current is

$$I_e = \alpha_I I_c + I_{ebo} \tag{12-10}$$

The quantity

$$\alpha_I = (I_e - I_{ebo})/I_c \tag{12-11}$$

is known as the *inverse current gain*. Generally, $\alpha_I < \alpha$.

Using α and α_I, we may relate the reverse currents as follows

$$I_{cbo} = I_{cbs}(1 - \alpha\alpha_I) \tag{12-12}$$

$$I_{ebo} = I_{ebs}(1 - \alpha\alpha_I) \tag{12-13}$$

The a.c. (small-signal) operation of the transistor can be described in terms of the *dynamic* (*a.c.*) *current gain*

$$\alpha_d = \left.\frac{dI_c}{dI_e}\right|_{V_{cb}=\text{constant}} \tag{12-14}$$

The quantities α and α_d can easily be related by differentiating Eq. (12-8) with respect to I_e

$$\alpha_d = \alpha + I_e\, d\alpha/dI_e \tag{12-15}$$

In the active region, α is practically unaffected by variations in the emitter current at low injection levels. Then, we may write that

$$\alpha_d = (\text{approx.})\ \alpha \tag{12-16}$$

The d.c. current gain depends not only on the motion of carriers in the base and the probability of their recombination, but also on the processes in the emitter and collector junctions. Therefore, the coefficient α is usually represented as the product of the coefficients characterizing the processes in the base and the two junctions

$$\alpha = \gamma\alpha_h\alpha^* \tag{12-17}$$

Here,

$$\gamma = I_{ep}/I_e \tag{12-18}$$

is the injection ratio;

$$\alpha_h = I_{cp}/I_{ep} \tag{12-19}$$

is the transfer coefficient describing the transfer of holes through the base into the collector;

$$\alpha^* = I_c/I_{cp} \tag{12-20}$$

is the collector efficiency factor showing the contribution of the hole current to the total collector current. I_{ep} and I_{cp} in Eqs. (12-18) through (12-20) are the hole components of the emitter and the collector currents, respectively.

The injection ratio γ has already been discussed in Chapter 11. At $N_{ae} \gg N_{db}$, γ for transistors approaches unity.

The factor α_h obviously depends on the relation between the base width and the diffusion length. This quantity is defined by an equation given in reference [24]

$$\alpha_h = (\text{approx.})\, 1 - w^2/2L_p^2 \qquad (12\text{-}21)$$

As a rule, in transistors, $w/L_p \leqslant 0.2$ or 0.3 and, as a consequence, α_h = (approx.) 0.95 to 0.98.

The factor α^* may increase owing to the avalanche ionization in the collector junction. However, when V_{cb} is not very high, that is when the transistor operates under normal conditions far from the breakdown region, this effect is negligibly small ($M \approx 1$). The electron component of the collector current is also insignificant because n_p in the collector is small. So α^* may be deemed to be unity.

As a result, we may write

$$\alpha = (\text{approx.})\, \gamma\alpha_h \qquad (12\text{-}22)$$

In alloy transistors, the base cannot be made very thin ($w > > 20$ μm), so α mainly depends on α_h. In transistors with a thinner base, the controlling factor is the injection level γ.

In a transistor connected in a common-emitter circuit, the input current is I_b, and the output, I_c as in the common-base configuration. In this case, the performance of the transistor is expressed in terms of the *common-emitter d.c. current gain*, β (also known as the *beta current gain*). The expression for β may be derived by substituting Eq. (12-1) into Eq. (12-8) and solving it for I_c

$$I_c = \alpha I_b/(1-\alpha) + I_{cbo}/(1-\alpha) \qquad (12\text{-}23)$$

Let us re-write this expression thus

$$I_c = \beta I_b + I_{ceo} \qquad (12\text{-}24)$$

Here,

$$\beta = \alpha/(1-\alpha) \qquad (12\text{-}25)$$

and

$$I_{ceo} = I_{cbo}/(1-\alpha) = (1+\beta)\, I_{cbo} \qquad (12\text{-}26)$$

is the reverse collector current in the common-emitter configuration at $I_b = 0$.

Equation (12-25) relates α to β.

The expression for the beta gain can be derived by combining Eqs. (12-24) and (12-26)

$$\beta = (I_c - I_{cbo})/(I_b + I_{cbo}) \tag{12-27}$$

The a.c. (small-signal) operation of the transistor can be expressed in terms of the *dynamic (a.c.) beta current gain*

$$\beta_d = \frac{dI_c}{dI_b}\Big|_{V_{ce}=\text{constant}} \tag{12-28}$$

The d.c. and a.c. beta current gains are related as follows

$$\beta_d = \beta + (I_b + I_{cbo})\, d\beta/dI_b \tag{12-29}$$

The relation $\beta\,(I_b)$ is more pronounced than $\alpha(I_e)$. If this relation may be ignored without introducing large errors by assuming that $d\beta/dI_b = 0$, then

$$\beta_d = (\text{approx.})\ \beta \tag{12-30}$$

Because α = (approx.) 0.95 to 0.99, β reaches several tens or even hundreds.

It is appropriate to mention here that the physical processes being discussed characterize the transistor operation in the steady-state active region. Transients associated with rapid variations in voltage and current, or high-frequency voltages are not taken into consideration. In general, carriers moving in the base introduce time lag, so α and β are frequency dependent. This will be discussed in Sec. 12-7.

Thus, the currents in the transistor are characterized by a number of physical processes in the *P-N* junctions and the base, which are described in terms of appropriate parameters. These are very important when analysing the a.c. small-signal operation of the transistor. Most of them are differentials and used as what are known as *small-signal parameters* of the transistor, which will be discussed in Sec. 12-6.

Let us examine the main processes and parameters using as an example an alloy transistor operating in the active region.

Processes in the Emitter Junction. The forward voltage V_{eb} causes the injection of minority carriers across the emitter junction. This is characterized by the injection ratio γ [Eq. (11-14)] and the injection level δ [Eq. (10-32)]. The injection results in a current flowing across the junction, whose components have been discussed in Sec. 12-2.

An important parameter characterizing the relation $I_e = f\,(V_{eb})$ is the *dynamic resistance of the emitter junction*, $r_{e,d}$.

Putting V_{cb} = constant and neglecting for the present the effect that this voltage has on the processes in the emitter junction, we may write for the emitter current

$$I_e = I_{ebX}\,[\exp\,(eV_{eb}/kT) - 1] \tag{12-31}$$

or

$$V_{eb} = (kT/e) \log_e (I_e/I_{ebX} + 1) \qquad (12\text{-}32)$$

where I_{ebX} is the reverse emitter current at a specified reverse voltage V_{cb}.

The resistance $r_{e,d}$ may be determined by differentiating Eq. (12-32) in which the unity may be ignored because, when V_{eb} is not too low, $I_e > I_{ebX}$

$$r_{e,d} = \frac{\partial V_{eb}}{\partial I_e} = (\text{approx.}) \left.\frac{kT}{eI_e}\right|_{V_{cb}=\text{constant}} \qquad (12\text{-}33)$$

The *emitter junction capacitance*, C_e, as the junction capacitances in general, is composed of the barrier capacitance, $C_{e,b}$, and the diffusion capacitance, $C_{e,D}$. These have been discussed in Sec. 10-6. Using Eq. (10-70), we write for the diffusion emitter junction capacitance

$$C_{e,D} = (\text{approx.}) (e/kT) (w^2/2D_p) I_e \qquad (12\text{-}34)$$

The barrier capacitance may be defined with the help of Eq. (10-66)

$$C_{e,b} = (\text{approx.}) (\varepsilon\varepsilon_0 s_e/l_e) \sqrt{\varphi_c/(\varphi_c + |V_{eb}|)}$$

where l_e is the width of the barrier region in the emitter junction at $V_{eb} = 0$, and s_e is the emitter junction area.

Processes in the Base. As with the diode, the injection leads to excess holes in the base at the boundary of the emitter junction. The neutralization of the resultant hole space charge by electrons has been discussed in Sec. 11-2 in connection with a similar process in the crystal diode.

In contrast to the diode, the base in the transistor is sandwiched between the emitter and collector junctions. So, the concentration of minority carriers in the base varies differently from that in the thin-base diode, described by Eq. (10-35).

The manner in which the concentration of minority carriers occurs in the transistor base may be described by solving Eq. (9-113) on the assumption that the boundary conditions are established by the specified hole current $I_{ep} = \gamma I_e$ at the boundary of the emitter junction (at $x = 0$) and the reverse collector voltage V_{cb} (at $x = w$).

In this case, Eq. (9-113) is solved as follows

$$p(x) = \Delta p \sinh [(w-x)/L_p]/\cosh (w/L_p)$$
$$+ p_{n0} \{[\exp (e|V_{cb}|/kT) - 1] \cosh (x/L_p)/\cosh (w/L_p) + 1\} \qquad (12\text{-}35)$$

where

$$\Delta p = I_{ep} L_p/eD_p s_e \qquad (12\text{-}36)$$

If $V_{cb} < 0$, $|V_{cb}| \gg kT/e$ and $w \ll L_p$ which is true of the transistor operating in the active region, then Eq. (12-35) reduces to

$$p(x) = (I_{ep} w/eD_p s_e) (1 - x/w) \qquad (12\text{-}37)$$

Hence, in the active region the minority carriers are distributed in the transistor base according to a linear law.

Using the above relations, we may plot the distribution of minority carriers in the base for different regions (Fig. 12-5). In the figure, the extent of the base (0—w) bounded by the emitter and collector junctions is shown without taking into account variations in the barrier region width at different V_{eb} and V_{cb}. Referring to Fig. 12-5,

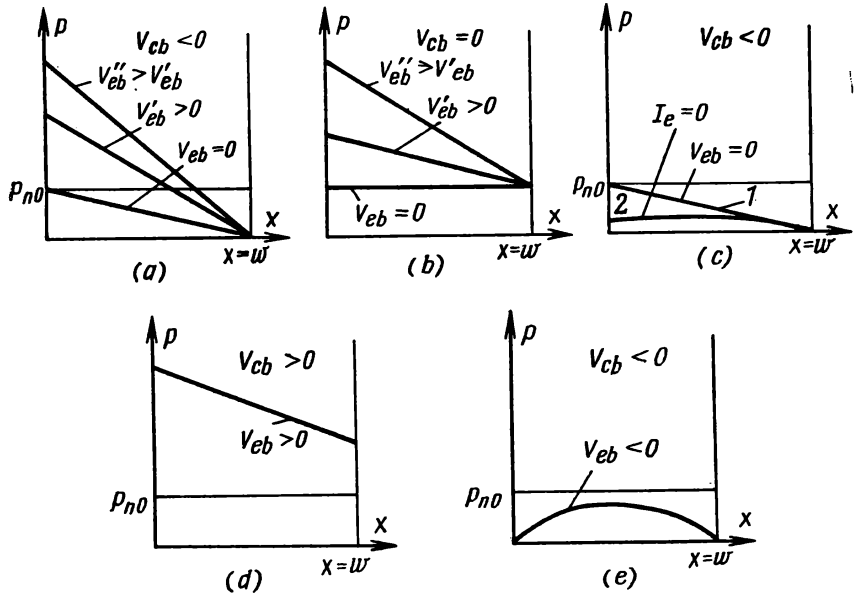


Fig. 12-5. Distribution of minority carriers (holes) in the transistor base (*a*) active region; (*b*) boundary between the active and saturated regions; (*c*) when measuring I_{cbs} (*1*) and I_{cbo} (*2*); (*d*) saturated region; (*e*) cut-off region

$p(w) = p_{n0}$ only at $V_{cb} = 0$; when $V_{cb} < 0$, $p(w) = 0$ due to the extraction of minority carriers; when $V_{cb} > 0$ (the saturation region), $p(w) > p_{n0}$. The plot of Fig. 12-5*c* illustrates the difference between I_{cbo} and I_{cbs}: at $V_{eb} = 0$, the hole concentration gradient in the base is higher than at $I_e = 0$ and, as a consequence, I_{cbs} is greater than I_{cbo}.

Differentiating Eq. (12-35) with respect to x, we may determine the hole concentration gradient in the base and, consequently, the diffusion current. In particular, we may get Eq. (12-21) for the coefficient α_h characterizing the transfer of holes through the base to the collector. The hole component of the collector current, I_{cp}, depends only on the first term in Eq. (12-35) because the second term does not involve I_{ep}. Differentiating Eq. (12-35) without the second term on the right-hand side and multiplying both sides of the new expression by $eD_p s_e$, we get an expression for I_{cp}, and,

dividing the same expression through by I_{ep}, we may write for the hole transfer coefficient

$$\alpha_h = \operatorname{sech}(w/L_p) \tag{12-38}$$

Because in the transistor $w \ll L_p$, we can expand sech (w/L_p) in a series and, limiting ourselves to the first two terms, get Eq. (12-21)

$$\alpha_h = (\text{approx.})\ 1 - w^2/2L_p^2$$

Base Spreading (or Bulk) Resistance. As a rule, the transistor base is made of a high-resistance material, so its spreading (or bulk) resistance r_b markedly affects the transistor operation. For example, at a high current, the voltage drop across r_b reduces the emitter bias. The base spreading resistance is mainly constituted by the resistance of the active base region, that is, the region sandwiched immediately between the emitter and collector junctions (see Fig. 12-1). The resistance of this region may be found by using a formula given in reference [24]

$$r_b = (\text{approx.})\ \rho_b/4\pi w \tag{12-39}$$

Because charges spread out as they move from the emitter to collector, the base resistance increases by 30 to 50% (due to the resistance offered by the passive base regions around the periphery of the emitter and collector junctions).

Base Width Modulation. In the active region, the collector junction is reverse-biased and its width depends on $|V_{cb}|$ according to Eq. (10-37). In an alloy transistor, the impurity concentration in the collector is much higher than in the base ($N_{ac} \gg N_{db}$). Also, in the active region, $|V_{cb}|$ is generally higher than φ_c, so Eq. (10-37) takes the form

$$l_c = (\text{approx.})\sqrt{2\varepsilon\varepsilon_0|V_{cb}|/eN_d} \tag{12-40}$$

As V_{cb} becomes more negative, the collector junction widens and the base grows narrower. This is known as the *base width modulation* (also called the *Early effect*).

Variations in the base width can easily be found by differentiating Eq. (12-40) with respect to V

$$|dw| = |dl_c| = \sqrt{(\varepsilon\varepsilon_0/2eN_d)}\,(dV_{cb}/\sqrt{|V_{cb}|}) \tag{12-41}$$

Variations in the base width considerably affect the physical processes in the base. A change in w brings about a change in the concentration gradient of minority carriers in the base (Fig. 12-6), which, at V_{cb} = constant, leads to an increase in the diffusion hole current density.

Voltage Feedback. This term is used to define the effect of V_{cb} on the above processes in the emitter junction. If an increase in $|V_{cb}|$ brings about a change of $dw = w - w'$ in the base width and an increase in the concentration gradient and the hole diffusion current, this means that more holes leave the emitter junction per unit time and I_e increases.

The emitter current can be restored to the original value by reducing the forward emitter-to-base voltage, V_{eb}. As a result, the hole concentration in the base will decrease to p_n' and the hole concentration gradient in the base will go down to the original value (Fig. 12-6).

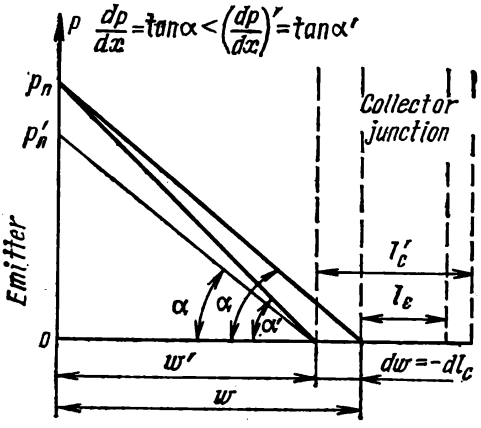


Fig. 12-6. The effect of base width modulation on the concentration gradient of minority carriers in the base

As is seen, I_e depends not only on V_{eb}, but also, to a certain extent, on V_{cb}. For the effects they produce on I_e, these voltages are compared in terms of the *voltage feedback factor*

$$\mu_{ce} = -\left.\frac{dV_{cb}}{dV_{eb}}\right|_{I_e=\text{constant}} \tag{12-42}$$

The relation between μ_{ce} and the physical parameters of the transistor can be established from the following considerations.

Variations in p_n with the applied voltage are described by Eq. (10-30)

$$p_n = p_{n0} \exp(eV/kT)$$

Differentiating this expression with respect to V and assuming that, in our case, $p_{n0} = p_n'$ and $V = V_{eb}$, we get

$$(p_n - p_n')/p_n = (e/kT)\, dV_{eb} \tag{12-43}$$

Referring to Fig. 12-6, the ratio on the left-hand side of Eq. (12-43) is equal to dw/w. Substituting dw/w into Eq. (12-43) and using Eq. (12-41), we obtain

$$\mu_{ce} = -\left.\frac{dV_{cb}}{dV_{eb}}\right|_{I_e=\text{constant}}$$

$$= \sqrt{2eN_d/\varepsilon\varepsilon_0}\,(ew\sqrt{V_{cb}}/kT) \text{ (at } I_e = \text{constant)} \tag{12-44}$$

The physical meaning of μ_{ce} is similar to that of the amplification factor μ of vacuum tubes. The only difference is that μ_{ce} defines the relative effect that V_{eb} (at the input) and V_{cb} (at the output)

have on the input current I_e, but not on the output current I_p, as is with vacuum tubes. If, however, we take into account that $I_c =$ $=$ (approx.) I_e, this difference will be insignificant.

As with the amplification factor of vacuum tubes, the minus sign in Eq. (12-42) shows that in order to maintain I_e constant, the increments dV_{eb} and dV_{cb} should be opposite in sign, that is, as V_{cb} becomes more negative, V_{eb} should decrease.

The factor μ_{ce} which is about 10^4 for transistors shows that the effect exerted by V_{cb} on I_e and, as a consequence, on I_c is insignificant.

Processes in the Collector Junction. According to Eq. (12-8), the collector current is

$$I_c = \alpha I_e + I_{cbo}$$

As already noted, all terms on the right-hand side of this equation depend on V_{cb}; I_{cbo}, though insignificantly, is affected by V_{cb} because variations in this voltage bring about variations in the width of the barrier region at the collector junction and, as a consequence, in I_c. According to Eq. (12-21), base width modulation causes α_h and I_e to change.

The relation $I_c = f(V_{cb})$ is closely associated with the *dynamic collector-junction resistance*

$$r_{c,d} = \frac{dV_{cb}}{dI_c}\bigg|_{I_e=\text{constant}} \qquad (12\text{-}45)$$

Its value as a function of the physical parameters, I_e and V_{cb_0}, can be determined by substituting Eq. (12-17) in Eq. (12-8), putting $\alpha^* = 1$, recalling Eqs. (12-21) and (12-41), and differentiating Eq. (12-8):

$$r_{c,d} = \sqrt{2eN_d|V_{cb}|/\varepsilon\varepsilon_0}\, L_p^2/\gamma w I_e \qquad (12\text{-}46)$$

Like the emitter-junction capacitance, the collector-junction capacitance, C_c, is composed of the barrier and diffusion capacitances. The barrier capacitance $C_{c,b}$ may be found by Eq. (10-66)

$$C_{c,b} = (\text{approx.})\ (\varepsilon\varepsilon_0 s_c/l_c) \times \sqrt{\varphi_c/(\varphi_c + |V_{cb}|)} \qquad (12\text{-}47)$$

where l_c is the width of the barrier region at the collector junction at $V_{cb} = 0$. The area s_c is taken to be equal only to the part of the collector junction area bounded by the effective base region, $s_c =$ (approx.) s_e, where s_e is the emitter junction area.

In most transistors, $C_{c,b}$ is relatively small, being from units to several tens of picofarads, but its reactance (finite at not too low frequencies) shunts the high-resistance collector junction, so the effect of $C_{c,b}$ may be rather substantial.

The diffusion capacitance of the collector junction, C_{cD}, defined as the ratio of an increment in the hole charge in the base to an

increment in V_{cb} applied to the junction, is important when the transistor operates in the inversion and saturation regions. In the active region, C_{cD} is much smaller than the diffusion capacitance of the emitter junction because variations in V_{cb} do not affect the charge of the injected carriers as they do in the emitter junction. The charge near the collector junction in the base varies only because the base width is modulated.

Thus, we have discussed the main physical processes in the transistor. The quantities characterizing these processes are often called the *physical parameters of the transistor*. They are used as elements in physical equivalent circuits for transistors. In addition, as will be shown later, there is a close connection between the physical parameters and the other transistor parameters characterizing d.c., small-, and large-signal operation, etc. This connection makes it possible to analyse the operation of transistors in circuits in the light of appropriate physical processes and develop electronic circuits intelligently.

12-3. THE TRANSISTOR AS A TWO-PORT

The transistor is an electric transducer the physical processes in which are utilized to convert the energy of external d.c. voltage sources into the signal energy. Generally, the voltages and currents in the transistor are related nonlinearly. So, the two-port equivalent to the transistor (Fig. 12-7) should be regarded as an active nonlinear two-port.

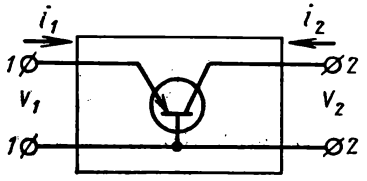


Fig. 12-7. Transistor as a two-port

As is known from network theory, the relations between the currents and voltages in a two-port may be represented by two functions of any two of the four arguments: i_1, i_2, v_1 and v_2. As a result, six pairs of functions are available. The performance of the transistor as a two-port is described in terms of only two of them:

$$\left.\begin{aligned} i_1 &= f_1(v_1;\ v_2) \\ i_2 &= f_2(v_1;\ v_2) \end{aligned}\right\} \tag{12-48}$$

$$\left.\begin{aligned} v_1 &= \varphi_1(i_1;\ v_2) \\ i_2 &= \varphi_2(i_1;\ v_2) \end{aligned}\right\} \tag{12-49}$$

These functions may be written as total differentials

$$\left.\begin{aligned} di_1 &= (\partial i_1/\partial v_1)\, dv_1 + (\partial i_1/\partial v_2)\, dv_2 \\ di_2 &= (\partial i_2/\partial v_1)\, dv_1 + (\partial i_2/\partial v_2)\, dv_2 \end{aligned}\right\} \tag{12-50}$$

$$\left.\begin{aligned} dv_1 &= (\partial v_1/\partial i_1)\, di_1 + (\partial v_1/\partial v_2)\, dv_2 \\ di_2 &= (\partial i_2/\partial i_1)\, di_1 + (\partial i_2/\partial v_2)\, dv_2 \end{aligned}\right\} \tag{12-51}$$

Y-(Admittance) Parameters. Equations (12-50) have already been used in Chapter 3 to examine the operation of the vacuum triode as a two-port. The coefficients of partial derivatives in Eqs. (12-50) have the dimension of admittance, so these equations may be written as*

$$\left.\begin{aligned} di_1 &= y_{11}\, dv_1 + y_{12}\, dv_2 \\ di_2 &= y_{21}\, dv_1 + y_{22}\, dv_2 \end{aligned}\right\} \tag{12-52}$$

The coefficients y_{11}, y_{12}, y_{21} and y_{22} make up the system of y-parameters of the transistor, which will be discussed in Sec. 12-6.

Hybrid (H-) Parameters. Most commonly, however, use is made of a system of equations using hybrid, or h, parameters, derived on the basis of Eqs. (12-51)

$$\left.\begin{aligned} dv_1 &= h_{11}\, di_1 + h_{12}\, dv_2 \\ di_2 &= h_{21}\, di_1 + h_{22}\, dv_2 \end{aligned}\right\} \tag{12-53}$$

In this system, two coefficients, h_{12} and h_{21}, have no dimension, h_{11} has the dimension of resistance, and h_{22} has the dimension of conductance. The physical meaning of these coefficients may be made out by putting in turn $di_1 = 0$ and $dv_2 = 0$ in Eqs. (12-53), which corresponds to the input of the two-port open-circuited and its output short-circuited

$$h_{11} = \frac{dv_1}{di_1}\bigg|_{dv_2=0} \tag{12-54}$$

is the input impedance;

$$h_{12} = \frac{dv_1}{dv_2}\bigg|_{di_1=0} \tag{12-55}$$

is the reverse transfer voltage ratio;

$$h_{21} = \frac{di_2}{di_1}\bigg|_{dv_2=0} \tag{12-56}$$

is the forward current transfer ratio;

$$h_{22} = \frac{di_2}{dv_2}\bigg|_{di_1=0} \tag{12-57}$$

is the output admittance.

As is seen, the system of equations utilizing h-parameters involves coefficients characterizing the most important physical processes in the transistor as a current-controlled device.

---

* The subscripts of the y-parameters correspond to the subscripts of the partial derivatives in Eqs. (12-50) as read from top to bottom.

In the active region, the emitter junction is forward-biased and, as a consequence, the input impedance is low, the collector junction is reverse-biased and the output impedance of the transistor is high. Thus, the h-parameters can most conveniently be measured with the input open-circuited and with the output shorted.

The Relation between the Parameters. Since the performance of transistors can be expressed in terms of both the h- and y-parameters, it is useful to know how they are related. For this purpose, it is convenient to write the equations for a two-port in matrix form. If, for example, we regard the currents and voltages as the rectangular components of some generalized vectors i and v, then the system of equations utilizing the y-parameters will be written in matrix form as follows

$$[i] = [y]\,[v] \tag{12-58}$$

where

$$[y] = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \tag{12-59}$$

is the admittance matrix of the two-port.

For the system using the h-parameters (generalized vectors w_1 and w_2) we may write

$$[w_1] = [h]\,[w_2] \tag{12-60}$$

where

$$[h] = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \tag{12-61}$$

The matrices $[y]$ and $[h]$ are related as follows

$$[y] = \frac{1}{h_{11}} \begin{bmatrix} 1 & -h_{12} \\ h_{21} & |h| \end{bmatrix} \tag{12-62}$$

$$[h] = \frac{1}{y_{11}} \begin{bmatrix} 1 & -y_{21} \\ y_{21} & |y| \end{bmatrix} \tag{12-63}$$

Here, $[y]$ and $[h]$ are the determinants of the respective matrices.

Table 12-1 gives the formulas relating the parameters of the two systems.

12-4. STATIC (D.C.) CHARACTERISTICS OF THE TRANSISTOR

As with vacuum tubes, the relations between the currents and voltages in transistors can be presented in the form of static (d.c.) characteristics. If, we choose as static characteristics relations between any two quantities out of the four on the condition that one

of the other two quantities is maintained constant, we will get twelve families of static characteristics for each of the three transistor circuit configurations. In practice, it would obviously be of little help to use so many interrelated curves. So, the relations between the currents and voltages in transistors are customarily represented by four families of characteristics most suitable for this purpose and connected with the most common *h*- arameters.

Among the static characteristics of the transistor most commonly used are

—the family of *input characteristics*:

$$V_1 = f_1(I_1) \text{ at } V_2 = \text{constant} \qquad (12\text{-}64)$$

—the family of *voltage feedback characteristics*:

$$V_1 = f_2(V_2) \text{ at } I_1 = \text{constant} \qquad (12\text{-}65)$$

—the family of *current transfer characteristics*:

$$I_2 = \varphi_1(I_1) \text{ at } V_2 = \text{constant} \qquad (12\text{-}66)$$

—and the family of *output characteristics*:

$$I_2 = \varphi_2(V_2) \text{ at } I_1 = \text{constant} \qquad (12\text{-}67)$$

The input and output currents and voltages, I_1, V_1 and I_2, V_2, respectively, vary in accordance with the transistor configuration.

Table 12-1

Unknown parameter	Known parameter	
	y	*h*
y	—	$y_{11} = 1/h_{11}$ $y_{12} = -h_{12}/h_{11}$ $y_{21} = h_{21}/h_{11}$ $y_{22} = (h_{11}h_{22} - h_{12}h_{21})/h_{11}$
h	$h_{11} = 1/y_{11}$ $h_{12} = -y_{12}/y_{11}$ $h_{21} = y_{21}/y_{11}$ $h_{22} = (y_{11}y_{22} - y_{12}y_{21})/y_{11}$	—

In practice, use is most often made of the input and output characteristics. The other two families are used but seldom and, unless they are specified in the data sheet, they may be deduced, if necessary, from the input and output characteristics.

Static Characteristics of an Idealized Transistor. Let us analyse the relations given by Eqs. (12-64) and (12-67) using, as an example,

a transistor connected in a common-base circuit. In the equivalent circuit of Fig. 12-8*a*, the *P-N* junctions are presented by two diodes connected in opposition. Owing to this arrangement, we may examine the operation of the transistor in the active and inversion regions. In the active region, the collection of the charge carriers injected across the emitter junction is accounted for by the current

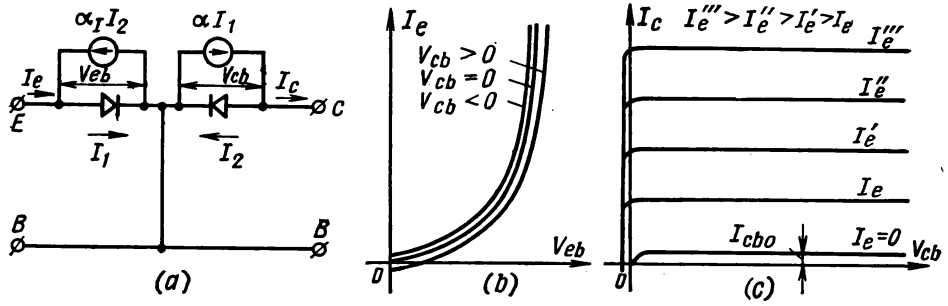


Fig. 12-8. (*a*) Transistor connected in a common-base circuit, (*b*) input characteristics and (*c*) output characteristics

generator αI_1. In the inversion region, charge carriers are injected across the collector junction and collected by the emitter. In the equivalent circuit, this is accounted for by the current generator $\alpha_I I_2$.

In this circuit, the base width modulation, the various bulk resistances, junction capacitances, etc. are ignored. So, the equivalent circuit of Fig. 12-8*a* represents an idealized transistor.

The injection current through any of the junctions can be found by Eq. (10-52) relating the current to the voltage across the respective *P-N* junction, with the other junction short-circuited. Accordingly, the current injected across the emitter junction with the collector shorted is

$$I_1 = I_{ebs} [\exp (eV_{eb}/kT) - 1] \tag{12-68}$$

With the emitter shorted, the current injected across the collector junction is

$$I_2 = I_{cbs} [\exp (eV_{cb}/kT) - 1] \tag{12-69}$$

The collector and emitter currents are the algebraic sums of the injection and collection currents

$$I_e = I_1 - \alpha_I I_2 \tag{12-70}$$

$$I_c = \alpha I_1 - I_2 \tag{12-71}$$

Substituting Eqs. (12-68) and (12-69) into the above expressions, we get

$$I_e = I_{ebs}\,[\exp(eV_{eb}/kT) - 1] - \alpha_I I_{cbs}\,[\exp(eV_{cb}/kT) - 1] \quad (12\text{-}72)$$

$$I_c = \alpha I_{ebs}\,[\exp(eV_{eb}/kT) - 1] - I_{cbs}\,[\exp(eV_{cb}/kT) - 1] \quad (12\text{-}73)$$

Solving Eq. (12-72) for V_{eb}, we get an expression for the input characteristics of an idealized transistor

$$V_{eb} = (kT/e)\log_e\{I_e/I_{ebs} + 1 + \alpha\,[\exp(eV_{cb}/kT) - 1]\} \quad (12\text{-}74)$$

Re-writing Eq. (12-72) for $[\exp(eV_{eb}/kT) - 1]$ and substituting it into Eq. (12-73), we get

$$I_c = \alpha I_e - I_{cbo}\,[\exp(eV_{cb}/kT) - 1] \quad (12\text{-}75)$$

In obtaining the solutions (12-74) and (12-75), use has been made of Eqs. (12-12) and (12-13) and also of the expression

$$\alpha I_{ebo} = \alpha_I I_{cbo} \quad (12\text{-}76)$$

interrelating the four independent parameters and derived by the authors of the equivalent circuit from a more detailed analysis.

Equations (12-74) and (12-75) describe the input and output characteristics of an idealized transistor (Fig. 12-8*b* and *c*).

When $V_{cb} = 0$, Eq. (12-74) takes the form

$$V_{eb} = (kT/e)\log_e(I_e/I_{ebs} + 1) \quad (12\text{-}77)$$

In the active region when $V_{cb} < 0$ and $|V_{cb}| \gg kT/e$, the term in the square brackets in Eq. (12-74) tends to -1. So, assuming that $1 - \alpha =$ (approx.) 0, we may re-write Eq. (12-74) as

$$V_{eb} = \text{(approx.)}\,(kT/e)\log_e(I_e/I_{ebs}) \quad (12\text{-}78)$$

As a consequence, in the active region ($V_{cb} < 0$), the characteristics should lie above the curve plotted at $V_{cb} = 0$.

In the saturation region ($V_{cb} > 0$), the characteristic shifts downward and to the right of the curve plotted at $V_{cb} = 0$.

In other words, at a specified V_{eb}, I_e is higher in the active region and lower in the saturation region than it is at $V_{cb} = 0$. This is because of the difference in minority-carrier concentration gradient between the two regions (see Fig. 12-5).

At $V_{cb} < 0$, the output characteristics correspond to the active region. In this case, at $|V_{cb}| \gg kT/e$, the term in the square brackets in Eq. (12-75) tends to -1, and this expression takes the form identical to Eq. (12-8)

$$I_c = \alpha I_e + I_{cbo} \quad (12\text{-}79)$$

At zero emitter current, the collector draws I_{cbo} whose value in the selected equivalent circuit of the transistor is unaffected by V_{cb} and depends only on the concentration of carriers in the base and collector. As the emitter current increases, I_c rises, too. Because

in the equivalent circuit α is independent of I_e, equal increments in I_e ($I_e''' - I_e'' = I_e'' - I_e' \ldots$) bring about equal increments in the collector current.

In the saturated region ($V_{cb} > 0$), holes are injected from the collector into the base, which is opposite to the holes moving from the emitter through the base to the collector. As this takes place, the collector current suddenly falls.

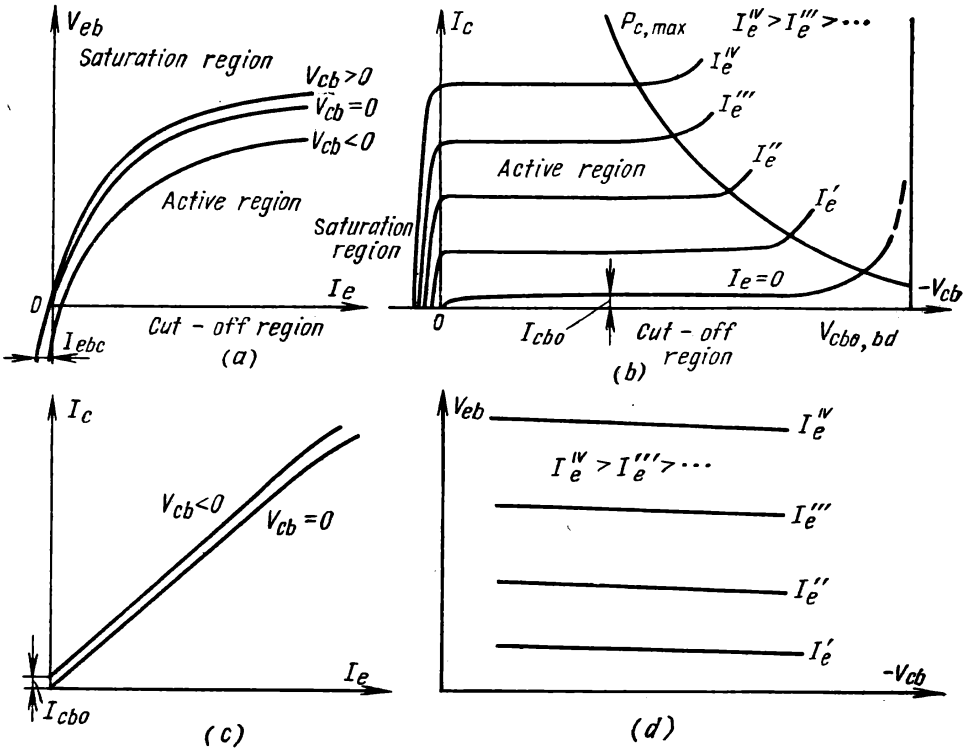


Fig. 12-9. Static characteristics of a transistor connected in a common-base circuit

(*a*) input; (*b*) output; (*c*) current transfer; (*d*) feedback

Because of some factors not taken into account in the adopted equivalent circuit, real transistor characteristics differ somewhat from the idealized characteristics, although, in general appearance, they are similar to those of Fig. 12-8*b* and *c*.

Static Characteristics of a Real Transistor Connected in a Common-Base Circuit. These are shown in Fig. 12-9*.

* Here and elsewhere, the voltage polarity on the characteristics is given with reference to a *P-N-P* transistor. For an *N-P-N* transistor, the polarity must be reversed. The positive directions of the currents are taken as follows: I_e, to the emitter in the external circuit; I_c, from the collector in the external circuit; I_b, from the base in the external circuit.

Input Characteristics. At $V_{cb} = 0$, the input characteristic, $V_{eb} = f_1(I_e)$ (Fig. 12-9*a*), is practically close to the volt-ampere characteristic of a real diode. The differences between the volt-ampere characteristics of a real and an idealized diode have been discussed in Sec. 11-3. At $V_{eb} < 0$, $I_e = I_{ebs}$. This current and its relation to V_{eb} depend on some physical parameters of the material (the relative magnitude of the thermal, generation and leakage currents), the junction area, and the like. At $V_{eb} > 0$, the characteristic differs from that of an idealized transistor due to recombination at the emitter junction and also because of the voltage drop across the base bulk resistance.

In the active region ($V_{cb} < 0$), the real characteristic shifts in the direction of high emitter currents much more than the characteristic of an idealized transistor. This can be attributed to the same factors as in the idealized transistor, and also to the base width modulation. As V_{cb} grows more negative, the collector junction becomes wider, the base length decreases and, at a constant V_{eb}, I_e increases because of the rising concentration gradient of minority carriers in the base (see Fig. 12-6).

In the saturation region ($V_{cb} > 0$), the shift of the real characteristic in the direction of low I_e is also affected by the base width modulation.

Output Characteristics. In the active region at negative V_{cb}, the output characteristics, $I_c = \varphi_2(V_{cb})$ (Fig. 12-9*b*), insignificantly differ from the characteristics of an idealized transistor. The difference is mainly due to the fact that I_c slightly increases as V_{cb} becomes more negative (that is, the characteristics are not parallel to the x-axis). This can be traced to the effect of base width modulation.

More specifically as V_{cb} grows more negative, the base becomes narrow, the hole concentration gradient in the base increases and so does I_e. In order to maintain I_e constant, it is necessary to reduce V_{eb}, taking care to keep dp/dx = constant (see Fig. 12-6). When I_e is held constant, but the base is reduced in width, the probability that holes will recombine in the base decreases, α increases and, as a consequence, so does I_c.

This effect may be allowed for by extending Eq. (12-79) to include an additional term

$$I_c = \alpha I_e + I_{cbo} + |V_{cb}|/\bar{r}_{c,d} \qquad (12\text{-}80)$$

where $\bar{r}_{c,d}$ is the averaged dynamic resistance of the collector junction defined by Eqs. (12-45) and (12-46).

Another difference is due to the fact that equal changes in the emitter current ($I_e''' - I_e'' = I_e'' - I_e' \ldots$) do not result in equal changes in I_c. This is because α decreases with increasing I_e for some reasons to be discussed later.

At very negative V_{cb}, the collector current rises since the collector junction approaches the breakdown region. As $|V_{cb}|$ increases, the multiplication factor M becomes considerably greater than unity [see Eq. (10-60)], and at $|V_{cb}| = V_{bd,av}$, it tends to infinity. An increase in M brings about an increase in I_c.

The voltage V_{cb} at which $I_c = MI_{cbo}$ ($I_e = 0$) tends to infinity is designated for transistors as $V_{cbo,bd}$. At high I_c, the avalanche breakdown may change to the thermal breakdown which will destroy the transistor.

The family of output characteristics also includes the curve of the maximum safe collector power dissipation, $P_{c,max}$, which bounds I_c and V_{cb} from above.

In the saturated region ($V_{cb} > 0$), the collector junction is rendered conducting, holes begin to move from the collector into the base, and I_c suddenly falls. In this region, the real characteristics insignificantly differ from the characteristics of an idealized transistor.

Current Transfer Characteristics. These characteristics, $I_c = \varphi_1(I_e)$, at V_{cb} = constant (Fig. 12-9*c*), are almost straight lines sloping to the x-axis at an angle smaller than $\pi/4$ because α = (approx.) 0.95 to 0.99. When V_{cb} is negative, the slope of the characteristic slightly increases because the collector junction expands and reduces the base width, so that α increases. As this takes place, the coefficient α_h increases according to Eq. (12-21). At $I_e = 0$, $I_c = I_{cbo}$, and the characteristic does not start from the origin.

As I_e increases, the characteristics somewhat depart from straight lines, and α decreases. This occurs because the carriers injected from the emitter into the base modulate the base resistance, and the injection ratio γ goes down.

Feedback Characteristics. These characteristics, $V_{eb} = f_2(V_{cb})$ at I_e = constant (Fig. 12-9*d*), show the relative effect of the emitter and collector voltages on the emitter current. Referring to the characteristics, V_{eb} markedly affects the height of the potential barrier at the emitter junction and the emitter current, whereas V_{cb} has a slight effect on I_e (due to base width modulation). This is reflected in the slight slope of the characteristic towards the x-axis. The point is that in order to balance out an increase in $|V_{cb}|$ and to maintain I_e constant, one must reduce V_{eb}. Adjacent curves are spaced unequal distances apart for equal changes in I_e because I_e rises exponentially with growing V_{eb}.

Static Characteristics of a Transistor Connected in a Common-Emitter Cirruit. The *input characteristics*, $V_{be} = f_3(I_b)$ (Fig. 12-10*a*), outwardly resemble the input characteristics of the transistor connected in a common-base circuit. The base current, however, is much smaller than the emitter current and so is its increment per

unit of V_{be}. The scale along the current axis in Fig. 12-10*a* is larger than it is in Fig. 12-9*a*. The effect that V_{ce} exerts on the curves is also different. As V_{ce} becomes more negative at V_{be} = constant, I_b decreases because the narrowing base naturally reduces the probability that minority carriers will recombine in the base.

At $V_{be} = 0$ and $V_{ce} < 0$, the base draws a current equal to the reverse emitter-collector current, $I_b = I_{ces}$, with the base short-

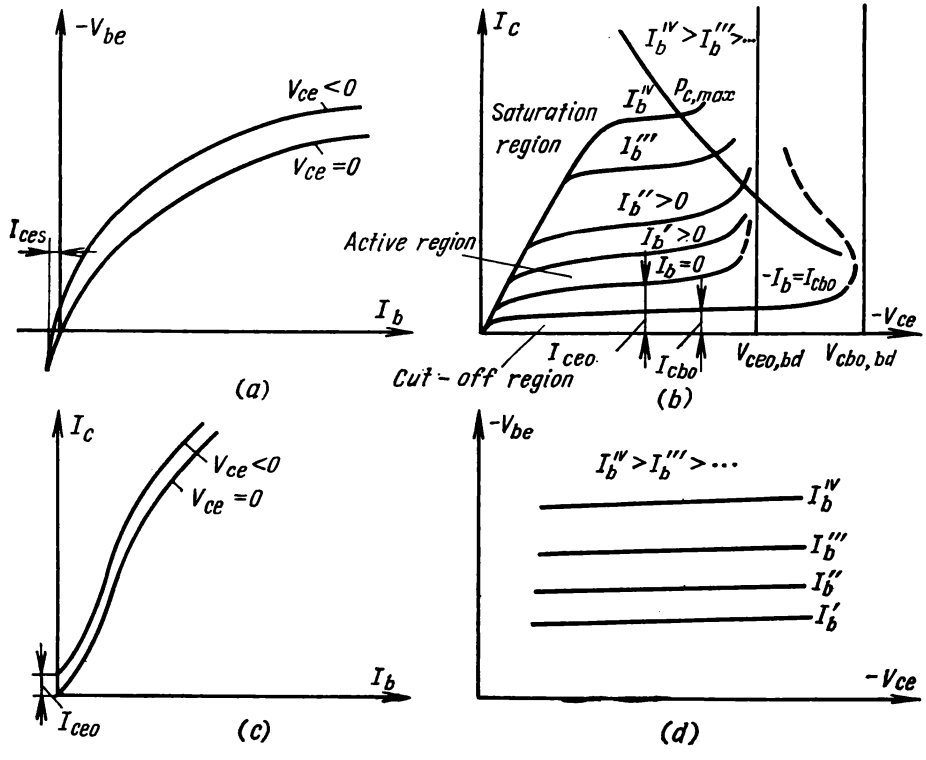


Fig. 12-10. Static characteristics of a transistor connected in a common-emitter circuit

(*a*) input; (*b*) output; (*c*) current transfer; (*d*) feedback

circuited. As has been agreed (see the footnote on page 302), the negative direction for the base current in the external circuit is towards the base. When V_{be} is positive, the negative base current slightly rises. Why this happens so will be discussed in connection with the output characteristics.

Output Characteristics. These characteristics, $I_c = \varphi_4 (V_{ce})$ at I_b = constant (Fig. 12-10*b*), differ from the respective characteristics for the common-base arrangement. When $V_{be} = 0$ and $V_{ce} = 0$, the potential difference between any pair of electrodes is zero and

there are no currents flowing in the transistor. If we maintain $V_{be} = 0$ (the base is short-circuited, Fig. 12-11), but apply to the collector a voltage, $-V_{ce}$, negative with respect to the emitter, the collector circuit will draw a reverse current $I_{ces} = I_{cbs}$ because, when the emitter is shorted, the common-base configuration no longer differs from the common-emitter configuration. In the base, I_{ces} is a sum of two components, namely the current flowing from the emitter junction to the collector junction due to the hole concentration gradient in the base (see Fig. 12-5*c*, curve *1*) and $-I_b$ entering the base from the shorted emitter. The latter depends only on V_{ce} and the collector junction resistance. If we open-circuit the emitter, the base current will remain unchanged, the first component of the collector current will be zero, the hole distribution in the base will correspond to curve *2* in Fig. 12-5*c*, and the circuit itself will be identical to the common-base configuration with the emitter open-circuited. As a consequence, the collector current will go down to I_{cbo}, and $-I_b = I_{cbo}$ since the current flows in a series circuit. This case ($V_{eb} = 0$) occurs at the boundary between the cut-off and active regions (the lower curve in Fig. 12-10*b*).

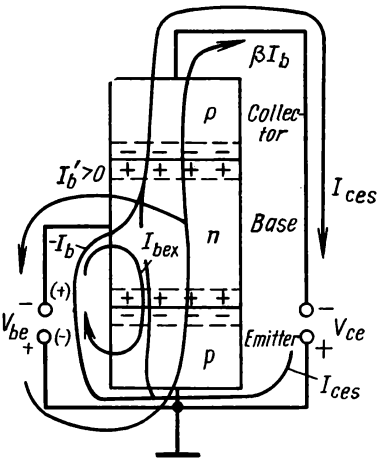


Fig. 12-11. Currents in a transistor connected in a common-emitter circuit

If the base is supplied with a positive voltage ($V_{be} > 0$), the emitter junction will be reverse-biased, and a reverse base current, I_{bex}, will flow in the base (Fig. 12-11). Some holes (minority carriers in the base), which would leave the base across the collector junction at $V_{be} = 0$, will now flow across the emitter junction, and the collector current will decrease.

The part of the output characteristics lying below the $-I_b = I_{cbo}$ curve corresponds to the cut-off region.

When the base is supplied with a negative voltage ($V_{be} < 0$), the emitter junction is rendered conducting, and the transistor changes to the active region, provided $| V_{ce} | > | V_{be} |$ (the collector junction is cut-off). Otherwise ($| V_{ce} | < | V_{be} |$), the potential difference between the collector and base will be positive, the collector junction will be forward-biased and the transistor will operate in the saturation region.

Thus, the active region of operation is associated with that part of the characteristic where $| V_{ce} | > | V_{be} |$, rather than with the entire region above the curve plotted for $-I_b = I_{cbo}$. On the family

of output characteristics (Fig. 12-10*b*), the boundary between the saturated and active regions is distinct; it passes through the points of inflection of the curves. In the saturation region, I_c suddenly falls as V_{ce} becomes less negative; in the active region, the curves $I_c = \varphi_4 (V_{ce})$ are more gradual.

Let us first examine the characteristics in the active region. When a negative voltage (with respect to the emitter) is applied to the base, a current due to the injection of minority carriers into the base flows across the emitter junction. Most of these carriers pass through the base and are intercepted by the collector junction; the others experience recombination and constitute in the base circuit a positive current I_b' which flows out of the base and is subtracted from $-I_b$ entering the base (Fig. 12-11). As V_{be} becomes more negative, I_b' increases and, at a certain value of V_{be}, the resultant base current becomes zero ($I_b = 0$), which is equivalent to open-circuiting the base. As this takes place, the collector current increases because, in addition to I_{cbo}, it also absorbs the collection current βI_b. Since in this case $| I_b | = | I_{cbo} |$, the resultant collector current, designated I_{ceo}, is

$$I_{ceo} = I_{cbo} + \beta I_b = I_{cbo} (1 + \beta) \qquad (12\text{-}81)$$

This expression is identical with Eq. (12-26) derived earlier. This case is represented by the characteristic plotted for $I_b = 0$ in Fig. 12-10*b*.

As V_{be} becomes still more negative, the emitter current increases and, consequently, so do the positive base current I_b and the collection current βI_b. Owing to this, the curves for high I_b correspond to rising I_c. As I_b rises, however, equal increments in I_b ($I_b''' - I_b'' = I_b'' - I_b' \ldots$) bring about unequal increments, ΔI_c, in the collector current. This is because β depends on I_b; why this is so will be discussed later when examining the characteristic $I_c = \varphi_3 (I_b)$.

In the active region, the output characteristics of the common-emitter configuration differ from those of the common-base arrangement in that the output current, I_c, is more dependent on the output voltage, V_{ce}. This happens for the following reasons. As V_{ce} becomes more negative, the collector junction expands and the base grows narrow. This reduces the probability of recombination in the base and, as a consequence, the base current. In order to maintain I_b constant, V_{be} should be raised. As a result, the injection of carriers from the emitter into the base increases and, consequently, so does the collector current. However, I_b depends on V_{be} only slightly, so, to maintain I_b constant, V_{be} should be changed considerably. Thus, in a transistor connected in the common-emitter circuit, I_c increases with increasing V_{ce} because I_e has to be raised to maintain I_b constant.

In the active region, the relation $I_c = \varphi_4 (V_{ce})$ can be described analytically by Eq. (12-24) if, by analogy with Eq. (12-80), we extend it to include an additional term characterizing the relation between I_c and V_{ce}

$$I_c = \beta I_b + I_{ceo} + V_{ce}/\bar{r}^*_{c,d} \tag{12-82}$$

where

$$\bar{r}^*_{c,d} = \bar{r}_{c,d}/(1+\beta) \tag{12-83}$$

Now, let us examine the characteristics in the saturation region. As already noted, the saturation region is represented by that part of the output characteristics where I_c suddenly falls as V_{ce} becomes less negative. At the points of inflection, $|V_{ce}| = |V_{be}|$. As $|V_{ce}|$ decreases further, the potential difference between the collector and base becomes positive, the collector junction is forward-biased, and the flow of holes from the base into the collector is balanced out by the opposite diffusion of holes from the collector. The collector current rapidly falls with decreasing $|V_{ce}|$. The characteristics plotted for the lower I_b = constant lie in the region where V_{be} is less negative, so a transition to the saturation region occurs at lower values of V_{ce}.

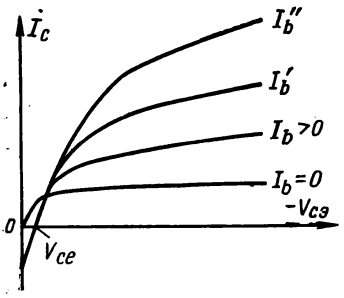


Fig. 12-12. Tail of the output characteristics of a common-emitter transistor

Figure 12-12 shows the tail of the output characteristics of the common-emitter arrangement plotted on an enlarged scale. The collector current is zero at a certain negative, collector-to-emitter voltage, V_{ce}. This voltage is low, being several tenths of a volt for silicon transistors. Using Eqs. (12-74) and (12-75) for an idealized transistor, we may show that V'_{ce} = (approx.) $(kT/e) \log_e \alpha_I$. An analysis of the physical factors involved lies outside the scope of this text. The fact itself, however, is important for micropower, especially integrated, transistors.

In the common-emitter circuit (where $I_b = 0$), as in the common-base configuration, a considerable increase in the collector voltage may lead to breakdown. However, the voltage, $V_{ceo,\,bd}$, at which breakdown now occurs is smaller than $V_{cbo,\,bd}$ for the common-base arrangement. The point is that the avalanche multiplication of carriers at the collector junction causes the alpha gain to increase to about M times its original value, where M is the multiplication factor defined by Eq. (10-59).

So

$$\beta = M\alpha/(1 - M\alpha) \tag{12-84}$$

and the avalanche breakdown ($\beta \to \infty$) occurs at $M\alpha \to 1$.

We can find $V_{ceo,\,bd}$ by substituting $M = 1/\alpha$ into Eq. (10-60)

$$V_{ceo,\,bd} = V_{cbo,\,bd} \sqrt[b]{1-\alpha} \qquad (12\text{-}85)$$

At $b = 3$ and $\alpha =$ (approx.) 0.98, the breakdown voltage for a transistor connected in a common-emitter circuit is one third of $V_{cbo,\,bd}$.

In the common-emitter configuration, the breakdown voltage depends on I_b. As the base current becomes more positive, the breakdown voltage decreases; in contrast, at $I_b < 0$, breakdown occurs at $|V_{ce}| > |V_{ceo,\,bd}|$.

Breakdown in transistors markedly depends on the value of the resistors connected in the base and collector circuits, R_b and R_c, respectively. When $R_b = 0$, breakdown occurs at voltages equal to $V_{cbo,\,bd}$; as R_b increases, the breakdown voltage decreases, so, when $R_b \to \infty$ (the base is open-circuited), it is equal to $V_{ceo,\,bd}$. This is why to avoid breakdown when a transistor is connected in a circuit, the base should be connected first.

The function of R_c is to limit the collector current. Should R_c be decreased to a certain value, the avalanche breakdown may change to the irreversible thermal breakdown.

In transistors with a high base resistivity, what is known as the *punch-through* takes place when, at very negative V_{ce}, the part of the collector junction lying in the base expands so that its space charge region touches the respective region of the emitter junction. The base region ceases to exist ($w = 0$), and, according to Eq. (12-21), $\alpha_h = 1$, the beta gain tends to infinity, and breakdown occurs.

Current Transfer Characteristics. These characteristics, $I_c = \varphi_3(I_b)$ (Fig. 12-10*c*), make a much greater angle with the input current axis than they do in the common-base configuration (in Fig. 12-10*c*, the scale along the I_b-axis is larger than that along the I_e-axis in Fig. 12-9*c*). This is quite natural, because I_b accounts for but an insignificant part (0.01 to 0.05) of the emitter current and, as a consequence, is a small fraction of I_c. The departure of $\beta(I_b)$ from a straight line is for the same reason as for $\alpha(I_e)$. However, when α changes by as little as 5% (from 0.9 to 0.95), β more than doubles. So on the $I_c = \varphi_3(I_b)$ characteristic, β falls significantly not only in the region of high I_b, which corresponds to decreasing α at high I_e, but also at very low I_b. The alpha and beta gains decrease at low currents because the relative effect of the recombination current in the emitter junction becomes more pronounced and, as a consequence, the injection ratio γ goes down. Because of these variations in $\beta(I_b)$, the spacing between the output characteristics is anything but constant (Fig. 12-10*b*).

The shift of the characteristic at $V_{ce} < 0$ toward the collector current axis is due to an increase in I_e and, consequently, I_c at $I_b =$ constant (this effect was discussed in more detail in connection

with the output characteristics). As already noted, at $I_b = 0$ and $V_{ce} < 0$, the collector current is equal to I_{ceo}.

Feedback Characteristics. These characteristics, $V_{be} = f_4(V_{ce})$ at I_b = constant (Fig. 12-10*d*), differ from those for the common-base configuration in their slope. We have already touched upon the cause for this difference in the discussion of the effect of V_{ce} on the input characteristics. The point is that I_b decreases as $|V_{ce}|$ increases due to base width modulation, in contrast to the typical rise in I_e with rising V_{cb} in the common-base configuration.

12-5. STATIC (D.C.) TRANSISTOR PARAMETERS

As their name implies, the static parameters of a transistor characterize its operation when only d.c. sources are connected to its terminals.

Static parameters are chosen so that differences in the static characteristics of a transistor in various regions of operation can be described with a minimum number of such parameters. Accordingly, we can use parameters for the cut-off, active and saturation regions of operation. Static parameters also include those describing the characteristics near the breakdown region.

Static Parameters in the Active Region. The performance of a transistor in the active region is expressed in terms of the *common-emitter forward current transfer ratio*

$$h_{21e} = (I_c - I_{cbo})/(I_b + I_{cbo}) \qquad (12\text{-}86)$$

This expression agrees with the already familiar Eq. (12-27). So, h_{21} is, in fact, the beta current gain, β.

In most cases, however, this parameter is defined as

$$h_{21e} = I_c/I_b \qquad (12\text{-}87)$$

with I_{cbo} neglected which is legitimate, provided that $I_b \geqslant 20I_{cbo}$. This inequality holds for most transistor applications except operation with extremely small signals and at high temperatures.

As a rule, the dependence of I_c on V_{cb} and V_{ce} is neglected when designing electronic circuits. In the circumstances, the parameter h_{21e} uniquely locates the quiescent point on the static characteristics of the transistor, provided that I_b or I_e is specified.

Another parameter used to describe the performance of a transistor in the active region is the *common-emitter forward transfer admittance*

$$Y_{21e} = \frac{I_c}{V_{be}}\bigg|_{V_{ce}=\text{constant}} \qquad (12.88)$$

Static Parameters in the Cut-Off Region. These parameters determine the thermal instability of transistors, and they are used in

all circuit calculations. They include the following currents already discussed in the previous sections:

—*reverse (leakage or cutoff) collector current*, I_{cbo}, measured at a specified collector-to-emitter voltage and with the emitter open-circuited;

—*reverse (leakage or cutoff) emitter current*, I_{ebo}, measured at a specified emitter-to-base voltage and with the collector open-circuited;

—*reverse (leakage or cutoff) collector current*, I_{cbs}, measured at a specified reverse collector-to-base voltage and with the base shorted;

—*reverse (leakage or cutoff) emitter current*, I_{ebs}, measured at a specified reverse emitter-to-base voltage and with the base shorted;

—*reverse collector-to-emitter current*, measured at a specified reverse collector-to-emitter voltage V_{ce}. This current is designated I_{ceo} when the base is open-circuited; I_{ces} when the base is shorted; I_{ceR} when measured at a specified base-emitter resistance; I_{ceX} when measured at a specified reverse base-to-emitter voltage V_{be}.

The physical meaning of the reverse currents used as parameters has been discussed in Sec. 12-2. They are interrelated by Eqs. (12-12) and (12-13).

Static Parameters in the Saturated Region. The performance of a transistor in this region is described in terms of the voltages at the electrodes of this transistor connected in a common-emitter circuit:

—*collector-to-emitter saturation voltage*, $V_{ce,sat}$, measured at specified base and collector currents;

—*base-to-emitter saturation voltage*, $V_{be,sat}$, measured at specified base and collector currents.

When measuring $V_{ce,sat}$, and $V_{be,sat}$, the collector current is most often deemed to be equal to nominal, and the base current is taken according to the following relation:

$$I_b = K_{sat} I_b' \tag{12.89}$$

where K_{sat} is the *saturation factor*, and I_b' is the current at the boundary of saturation region.

Sometimes, the operation of a transistor is expressed in terms of the *saturation resistance*

$$r_{sat} = V_{ce\ sat}/I_{c\ sat} \tag{12-90}$$

where $I_{c,sat}$ is the direct collector current in the saturation region.

Static Parameters in the Breakdown Region. These are already familiar quantities, namely:

—*collector-to-base breakdown voltage*, $V_{cbo,bd}$, measured at a specified reverse collector current, I_{cbo}, and for $I_e = 0$;

—*collector-to-emitter breakdown voltage*, $V_{ceo,bd}$, measured at a specified current I_c.

It is defined by Eq. (12-85)

$$V_{ceo,bd} = V_{cbo,bd} \sqrt[b]{1-\alpha}$$

12-6. SMALL-SIGNAL OPERATION OF A TRANSISTOR

As with vacuum tubes, when a transistor is connected in a circuit, its circuits contain not only d.c. voltage sources but also input signal sources and load (resistors, transformers, etc.).

The signals handled by transistor circuits may have different waveforms (harmonic, modulated in a complex way, pulse, and so on) and frequencies.

One of the simplest cases is the use of a transistor as a sinusoidal small-signal amplifier operating at frequencies not over several

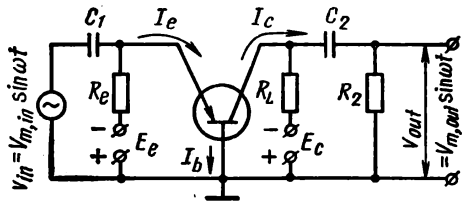


Fig. 12-13. Simplified amplifier circuit based on a transistor connected in a common-base circuit

MHz. A small signal is, for example, a sinusoidal voltage $v_s = V_{m,s} \sin \omega t$ whose amplitude, $V_{m,s}$, is such that, when halved or doubled, it will not affect the static characteristic being measured beyond the normal accuracy of the measurement of that characteristic, and the transistor itself may be regarded as a linear two-port.

The operation of a transistor in circuits handling higher frequency signals, pulses with short rise and fall times or large signals has a number of distinctions in each particular case, which will be discussed later.

The Transistor as a Small-Signal Amplifier. A simplified circuit diagram of an amplifier stage built around a transistor connected in a common-base circuit is shown in Fig. 12-13. The input circuit of the stage contains resistor, R_e, which sets the direct emitter current; the emitter circuit is supplied by a direct voltage source, E_e. Also, the input circuit contains the input signal source. The capacitor C_1 isolates the input signal circuit from the emitter-supply circuit. The collector is supplied by a direct voltage source, E_c, through a load resistor, R_L.

The Load Line. This is constructed in much the same manner as for a vacuum-triode amplifier (Sec. 3-4).

By Kirchhoff's voltage law, we may write for the collector (output) circuit

$$E_c = I_c R_L + V_{cb} \tag{12-91}$$

or

$$I_c = E_c/R_L - V_{cb}/R_L \tag{12-92}$$

When plotted in the output coordinates, I_c and V_{cb}, Eq. (12-92) yields a straight line which can easily be constructed if we know

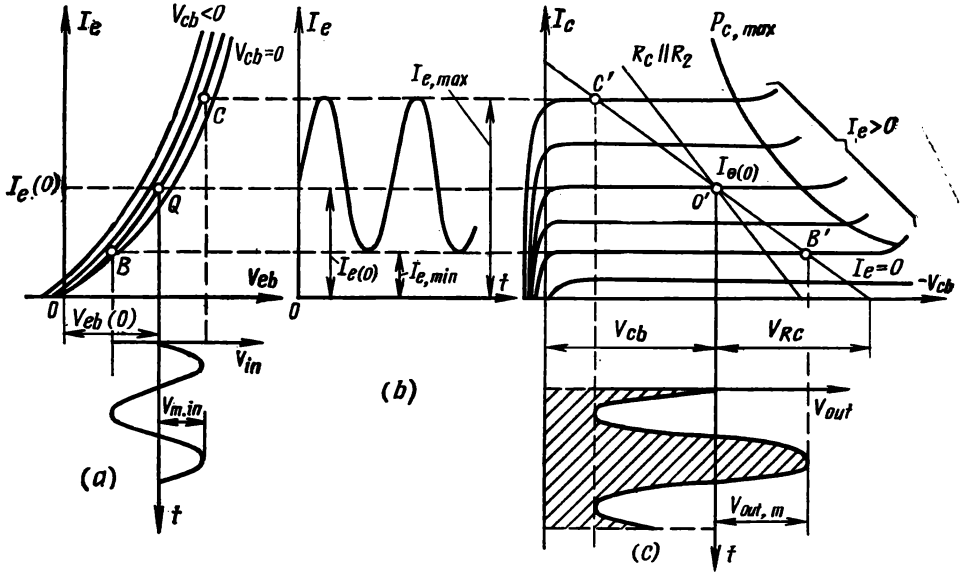


Fig. 12-14. Operation of the transistor as an amplifier

(*a*) input characteristics; (*b*) I_e waveforms; (*c*) construction of the load line on the family of output characteristics

the intercepts made by this line on the axes: at $I_c = 0$, $V_c = E_c$, and, at $V_{cb} = 0$, $I_c = E_c/R_L$ (Fig. 12-14*c*). Since I_e is set by the source E_e and resistor R_e, the operating point Q' is uniquely positioned on the output characteristics at the point where the load line intersects the output characteristic corresponding to the specified current $I_e = I_{e(0)}$. As with the vacuum-triode amplifier, E_c and R_L are constant at any point on the load line. By transferring the points where the load line intersects the output static characteristics on to the family of input static characteristics, we may construct the load line. As will be recalled, however, V_{cb} has almost no effect on I_e, so the input characteristics for different V_{cb} form a narrow bundle of curves spaced closely apart (Fig. 12-14*a*). There-

fore, as with the pentode, transistor data sheets and manuals usually specify only one input characteristic corresponding to the nominal V_{cb}.

From the foregoing it follows that the at-load (a.c.) input characteristic can be replaced with sufficient accuracy by the static (d.c.) input characteristic, as is done in Fig. 12-14*a*.

Principle of Operation of the Amplifier. This can be inferred from Fig. 12-14. When no signal is applied to the amplifier ($v_{in} = 0$), the emitter-to-base voltage, $V_{eb(0)}$, and the emitter current, $I_{e(0)}$, are set, that is, the Q-point in Fig. 12-14*a* is positioned by the supply source E_e and resistor R_e. When a signal is applied, the sinusoidal voltage, shown in Fig. 12-14*a* as the waveform v_{in} (t), is algebraically added to $V_{eb(0)}$, and the Q-point moves on the input characteristic along the path between points B and C which represent the peaks of v_{in}. If the signal is small, the portion of the input characteristic between points B and C may safely be replaced by a straight-line segment.

The emitter current varies in phase with the instantaneous value of $V_{eb(0)} + v_{in}$ (Fig. 12-14*b*). Since $I_e =$ (approx.) I_c ($\alpha \approx 0.98$ and more), variations in the output (collector) current, I_c, follow variations in the input (emitter) current, I_e. As this takes place, the Q-point moves on the output characteristics along the load line between points B' and C'.

As it flows through R_L, the collector current gives rise to a voltage drop across this resistor. So long as I_c increases (during the positive half-cycles of v_{in}), this voltage drop, $V_{R_L} = I_c R_L$, rises and the collector-to-base voltage falls according to Eq. (12-91). Conversely, so long as I_c decreases, V_{cb} increases. The collector-to-base voltage is the output voltage, so v_{out} varies in anti-phase with I_c and, as a consequence, with v_{in}.

As V_{cb} varies in anti-phase, it gives rise to what is known as the output circuit reaction; namely as v_{in}, I_e and I_c increase, V_{cb} decreases and tends to oppose the output current, I_c. However, I_c is very little affected by V_{cb} (the output characteristics are almost parallel to the x-axis), so this reaction, as in the pentode, is of minor significance.

Thus, if the collector circuit contains a high-value load resistor, the alternating voltage across this resistor ($V_{m,\,out}$ in Fig. 12-14*c*) will be many times the input voltage, $V_{m,\,in}$, and this is amplification. The amount of amplification is stated in terms of the voltage stage gain:

$$K_v = V_{m,\,out} / V_{m,\,in} \qquad (12\text{-}93)$$

In transistors used as amplifiers, as in other electronic devices, the output current should easily be controlled by the input signal

(the input current) and the output circuit reaction should be minimal.

In contrast to vacuum tubes, where the input resistance approaches infinity when the control grid draws no current, the input resistance of the transistor, $r_{e,\,d}$, is sufficiently small. So, when a second amplifier stage is coupled to the first, one must take into account the effect that its input resistance may have on the output circuit of the first stage. In the circuit of Fig. 12-13, this effect is taken care of by the resistor R_2 equal to the input resistance of the succeeding stage. The path for direct current from the output of the first stage to the input of the next is blocked by a capacitor, C_2. The path for alternating current, however, is completed by a resistor, R_2, which bypasses the load resistor R_L. Because of this, the slope of the load line changes ($R_L \parallel R_2$ in Fig. 12-14*c*).

We have discussed the circuit and operation of a transistor amplifier in a very simplfied way. For more detailed information of this matter, the reader should turn to a course on amplifiers.

Small-Signal Parameters. The small-signal operation of transistors is expressed, and the design and calculation of respective devices are carried out, in terms of what are known as the *small-signal parameters.* These are the coefficients of Eqs. (12-52) and (12-53) written for the linear two-port discussed in Sec. 12-3. The small-signal parameters are differentials (dynamic quantities).

As already noted, in the small-signal operation of a transistor, the path of its Q-point may safely be replaced by a straight-line segment. In this case, the transistor is equivalent to a linear two-port, and the differentials in the expressions for the y- and h-parameters may be replaced by the finite changes in its direct voltages and currents or by the peak values of the alternating voltages and currents.

As already noted, the h-parameters are used most of all. The physical meaning of the h-parameters as applied to a linear two-port has briefly been discussed in Sec. 12-3 in Eqs. (12-54) through (12-57). With reference to the transistor, the subscripts of the h-parameters are supplemented by letters designating the grounded (common) electrode, and the parameters themselves are represented by the ratios between the finite changes in voltages and currents.

H-Parameters for the Common-Base Configuration. Let us discuss these parameters using Eqs. (12-54) through (12-57) and recalling that for the common-base configuration $i_1 = I_e$, $i_2 = I_c$, $v_1 = V_{eb}$ and $v_2 = V_{cb}$. We shall also keep in mind that the h-parameters are related to the physical parameters of the transistor.

The *input impedance of the transistor*

$$h_{11b} = \frac{\Delta V_{eb}}{\Delta I_e}\bigg|_{V_{cb} = \text{constant}} \qquad (12\text{-}94)$$

is the ratio of an incremental change in the input voltage to the incremental change in the input current that caused it, with the transistor output shorted.

The relation between $h_{11b} = r_{e,d}$ and I_e is defined by Eq. (12-33). When I_e ranges between about 0.1 and 0.2 mA, h_{11b} is equal to a few kilohms, and at I_e = (approx.) 1 mA, it is equal to several tens of ohms.

The *voltage feedback ratio*

$$h_{12b} = \frac{\Delta V_{eb}}{\Delta V_{cb}} \Big| I_e = \text{constant} \tag{12-95}$$

is the ratio of an incremental change in the input voltage to the incremental change in the input current that caused it, when no a.c. signal is applied to the input circuit. In other words, h_{12b} defines the relative effect exerted by the input and output voltages on the input current.

The relation between $h_{12b} = -1/\mu_{ce}$ and the physical parameters and V_{cb} is described by Eq. (12-44). The value of h_{12b} ranges between about 10^{-3} and 10^{-4}.

The *forward current transfer ratio*

$$h_{21b} = \frac{\Delta I_c}{\Delta I_e} \Big| V_{cb} = \text{constant} \tag{12-96}$$

is the ratio of an increment in the output current to the increment in the input current that caused it, with the output shorted.

This parameter is the same as α_d. The physical processes in the transistor, affecting the value of h_{21b}, have been discussed in detail in Sec. 12-2, and the relation between α and I_e and V_{cb} has been considered in Sec. 12-4.

For most transistors, h_{21b} = 0.95 to 0.99.

The *output admittance*

$$h_{22b} = \frac{\Delta I_c}{\Delta V_{cb}} \Big| I_e = \text{constant} \tag{12-97}$$

is the ratio of an increment in the output current to the increment in the output voltage that caused it, with no a.c. signal in the input circuit.

The relation between $h_{22b} = 1/r_{c,d}$ and the physical parameters, I_e and V_{cb} is described by Eq. (12-46).

The output admittance is anywhere between 10^{-6} and 10^{-7} S.

H-Parameters for the Common-Emitter Configuration. In this arrangement, the input and output currents and voltages, $i_1 = I_b$, $i_2 = I_c$, $v_1 = V_{be}$ and $v_2 = V_{ce}$, are other than those in the common-base configuration, but the definitions given for the common-base circuit hold for this arrangement as well.

The *input impedance*

$$h_{11e} = \frac{\Delta V_{be}}{\Delta I_b}\Big|_{V_{ce}=\text{constant}} \tag{12-98}$$

differs in magnitude from h_{11b} because, in contrast to Eq. (12-94) the denominator in Eq. (12-98) is ΔI_b, the increment in the base current, which is considerably smaller than ΔI_e for the same ΔV_{eb}. In discussing the input characteristic of the common-emitter configuration, we have mentioned that the $I_b = f(V_{be})$ curves are more gradual which is an indication that in this arrangement the input impedance of the transistor is higher. The input impedance h_{11e} is $(1 + \beta)$ times h_{11b}, being several hundred ohms in the initial portion and a few kilohms at I_c = (approx.) 1 mA.

The *voltage feedback ratio*

$$h_{12e} = \frac{\Delta V_{be}}{\Delta V_{ce}}\Big|_{I_b=\text{constant}} \tag{12-99}$$

is identical to h_{12b}. Its value is about 10^{-3}.

The *forward current transfer ratio*

$$h_{21e} = \frac{\Delta I_c}{\Delta I_b}\Big|_{V_{ce}=\text{constant}} \tag{12-100}$$

is one of the most important parameters characterizing the performance of the transistor connected in a common-emitter circuit: h_{21e} = (approx.) β_d. The relation between this parameter and α_d and the physical processes affecting its value have already been discussed in Secs. 12-2 and 12-4.

Since α_d is usually between 0.95 and 0.99 or more, h_{21e} ranges between several tens or even hundreds.

The *output admittance*

$$h_{22e} = \frac{\Delta I_c}{\Delta V_{ce}}\Big|_{I_b=\text{constant}} \tag{12-101}$$

has the same meaning as h_{22b}: $1/h_{22b} = r^*_{c,d}$, where $r^*_{c,d}$ is the dynamic resistance of the collector junction in the common-emitter arrangement. Since the dependence of I_c on the output voltage is more pronounced than in the common-base configuration, h_{22e} is several tens of times h_{22b}, being about 5×10^{-5} S.

H-Parameters for the Common-Collector Configuration. For this configuration, $i_1 = I_b$, $i_2 = I_e$, $v_1 = V_{bc}$ and $v_2 = V_{ec}$.

The *input impedance* is defined as

$$h_{11c} = \frac{\Delta V_{bc}}{\Delta I_b}\Big|_{V_{ec}=\text{constant}} \tag{12-102}$$

It is approximately equal to h_{11e}.

The *voltage feedback ratio* is defined as

$$h_{12c} = \frac{\Delta V_{bc}}{\Delta V_{ec}} \Big| I_b = \text{constant} \qquad (12\text{-}103)$$

For this configuration, $h_{12c} = 1$.

The *forward current transfer ratio* is defined as

$$h_{21c} = \frac{\Delta I_e}{\Delta I_b} \Big| V_{ec} = \text{constant} \qquad (12\text{-}104)$$

This parameter is approximately equal to that for the common-emitter arrangement, because I_e = (approx.) I_c and, as a consequence, ΔI_e = (approx.) ΔI_c.

The *output admittance* is defined as

$$h_{22c} = \frac{\Delta I_e}{\Delta V_{ec}} \Big| I_b = \text{constant} \qquad (12\text{-}105)$$

Its value is the same as for the common-emitter configuration: h_{22c} = (approx.) 5×10^{-5} S.

The choice of expressions connecting the h-parameters for all three transistor configurations depends on the assumed directions of current flow in the two-port. Table 12-2 gives the formulas which apply when I_e and I_c enter the device. Here, $h_{21b} = -\alpha_d$.

Table 12-2

Common base	Common emitter	Common collector
h_{11b}	$h_{11e} \approx h_{11b}/(1+h_{21b})$	$h_{11c} \approx h_{11b}/(1+h_{21b})$
h_{12b}	$h_{12e} \approx h_{11b}h_{22b}/(1+h_{21b}) - h_{12b}$	$h_{12c} \approx 1$
h_{21b}	$h_{21e} \approx -h_{21b}/(1+h_{21b})$	$h_{21c} \approx -1/(1+h_{21b})$
h_{22b}	$h_{22e} \approx h_{22b}/(1+h_{21b})$	$h_{22c} \approx h_{22b}/(1+h_{21b})$

Y-Parameters. These are the coefficients of the partial derivatives in Eqs. (12-50) for a two-port. They have the dimension of conductance and are, in fact, the dynamic admittances of the transistor.

The y-parameters are used to describe the small-signal operation of transistors at high frequencies, when the currents and voltages in transistors are complex (see Sec. 12-7). This explains why in the small-signal operation the transistor admittances are used as y-parameters.

As with the h-parameters, the letters b, e and c in the subscripts of the y-parameters stand for the common-base, common-emitter and common-collector configurations, respectively.

The *input admittance*,

$$Y_{11} = \left.\frac{\Delta \dot{I}_1}{\Delta \dot{V}_1}\right| V_2 = \text{constant} \qquad (12\text{-}106)$$

is the ratio of an increment in the complex input current to the increment in the complex input voltage caused by the increment in current, with the output terminals short-circuited for alternating current.

The *reverse transfer admittance*

$$Y_{12} = \left.\frac{\Delta \dot{I}_1}{\Delta \dot{V}_2}\right| V_1 = \text{constant} \qquad (12\text{-}107)$$

is the ratio of an increment in the complex input current to the increment in the complex output voltage that caused it, with the input short-circuited for alternating current.

The *forward transfer admittance*

$$Y_{21} = \left.\frac{\Delta \dot{I}_2}{\Delta \dot{V}_1}\right| V_2 = \text{constant} \qquad (12\text{-}108)$$

is the ratio of an increment in the complex output current to the increment in the complex input voltage that caused it, with the output short-circuited for alternating current.

The *output admittance*

$$Y_{22} = \left.\frac{\Delta \dot{I}_2}{\Delta \dot{V}_2}\right| V_1 = \text{constant} \qquad (12\text{-}109)$$

is the ratio of an incremental change in the complex output current to the incremental change in the complex output voltage that caused it, with the input short-circuited for alternating current.

The physical meaning of the above admittances, currents and voltages for the common-base, common-emitter and common-collector configurations was defined in discussing the *h*-parameters.

Of the four *y*-parameters, only the absolute value of Y_{11}, which is the reciprocal of h_{11} (because in either case, the output is short-circuited for alternating current), has the same meaning as the respective quantity in the *h*-parameters and its physical equivalent, the dynamic resistance $r_{e,\,dyn}$ in the common-base configuration.

The absolute value of Y_{22} determined for V_1 = constant differs in the meaning from the output admittance h_{22} for which I_1 = = constant.

The formulas relating the *y*-parameters to the *h*-parameters are given in Table 12-1.

Determining Small-Signal Parameters from Static Characteristics. This can be carried out most easily for the *h*-parameters. All

the four parameters can be determined from two families of characteristics, the feedback and current transfer characteristics or the input and output characteristics (Fig. 12-15). To this end, we begin by meeting the two conditions for constancy of currents and voltages graphically. We specify the Q-point and draw through it lines parallel to the coordinate axes—these lines meet the first requirement for constancy of appropriate currents and voltages. The second requirement is met by the characteristic itself. For example, on the input characteristics plotted for the common-base configuration in Fig. 12-15*a*, we get I_e = constant and V_{cb} = constant. The

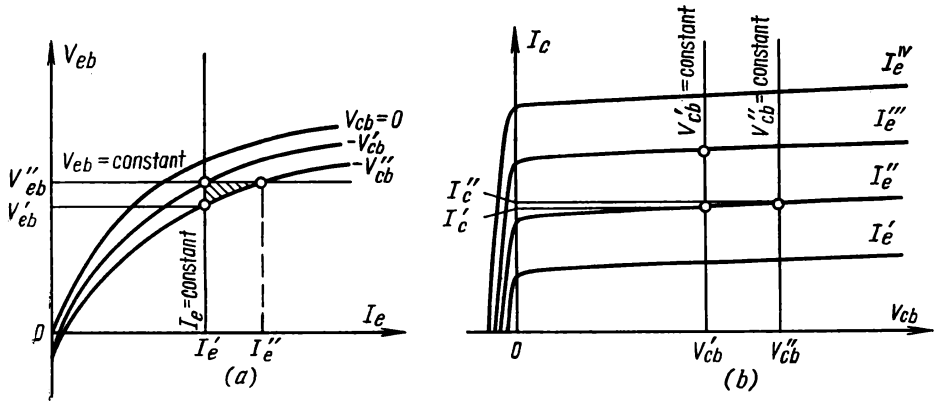


Fig. 12-15. Determining h-parameters from the static characteristics
(*a*) determining h_{11} and h_{12} from the family of input characteristics; (*b*) determining h_{21} and h_{22} from the family of output characteristics

points where the above lines intersect the two characteristics define the intercepts corresponding to the respective increments in voltages and currents. Thus, we get $\Delta I_e = I''_e - I'_e$, $\Delta V_{eb} = V''_{eb} - V'_{eb}$ and $\Delta V_{cb} = |V''_{cb}| - |V'_{cb}|$. The last increment is found as the difference in V_{cb} between two adjacent characteristic curves. The values of the h-parameters are found using Eqs. (12-94) through (12-97). Finally from the increments measured on the input characteristics, we find $h_{11} = \Delta V_{eb}/\Delta I_e$ and $h_{12} = \Delta V_{eb}/\Delta V_{cb}$.

It is obvious that the other two h-parameters can be found using the increments measured on the output characteristics of Fig. 12-15*b*. Similarly, we may determine the h-parameters from the families of the two other characteristics.

12-7. R. F. AND PULSE OPERATION OF THE TRANSISTOR

Up to now, we have discussed the transistor handling either direct voltages and currents or audio-frequency signals when the oscillation period, T, of the applied signal is much longer than the time

of the process in the transistor (carrier diffusion or drift time, charge or discharge time of the diffusion and barrier capacitances, etc.).

But, undoubtedly, the main physical processes occurring in the transistor in d.c. operation introduce a certain time lag, too. For one thing, there is a time lag associated with the diffusion of carriers in the base. For another thing, at radio frequencies, the performance of the transistor is affected by the emitter and collector capacitances, C_e and C_c. These capacitances are in parallel with the emitter and collector junction resistances $r_{e,j}$ and $r_{c,j}$. At audio frequencies, the reactances of C_e and C_c are very high, so their shunting effect may be neglected. As the frequency increases, the reactances decrease and their shunting effect becomes rather significant. Also, at radio frequencies, the charge and discharge time constants of C_e and C_c become comparable with the oscillation period.

In the general case, C_e and C_c include both the barrier and diffusion capacitances of the P-N junctions. The effect of the capacitances varies according as the junctions are forward- or reverse-biased.

Thus, when the transistor handles r.f. or pulse signals, the oscillation period or the rise and fall times of a voltage pulse may be so small that they are comparable with the duration of the physical processes in the transistor. This leads to phase shifts between voltages and currents in the device, the characteristic admittances become complex in nature, and the other parameters, notably the current gain, change in magnitude. As the frequency rises, the alpha current gain falls mainly due to a decrease in its components, the injection ratio γ and the transfer factor α_h characterizing the transfer of holes through the base.

If, for example, C_e is connected in parallel with the emitter junction resistance $r_{e,j}$, so the input admittance of the transistor is a complex quantity

$$Y_{11b} = 1/r_{e.j} + i\omega (C_{eD} + C_{e,b}) \qquad (12\text{-}110)$$

The capacitances C_{eD} and $C_{e,b}$ are both responsible for the time lag in the emitter junction, but they affect the operation of the junction differently. As will be recalled, C_{eD} depends on the charge ΔQ stored in the base due to the injection of minority carriers [see Eq. (10-68)]. So, the diffusion capacitance is charged, as it were, by minority carriers, and the time during which it charges is equal to the time necessary for the minority carriers in the base to attain a steady state distribution.

The barrier capacitance is mainly due to immobile charges constituted by ionized impurity atoms which remain unpaired in the barrier region of the junction as the majority carriers leave it. In other words, the barrier capacitance is charged by majority carriers, and its charge current has nothing to do with the injection and, consequently, is not transferred to the collector.

As follows from the foregoing, the injection current across the emitter junction is associated with the first two terms of rather than the entire Y_{11b} [see Eq. (12-110)]. Some of the current flowing across the emitter junction is due to the susceptance $i\omega C_{e,b}$. Because of this, the ratio of the injection current to the total current across the emitter junction decreases, and, as a consequence, so do the injection ratio γ and, according to Eq. (12-17), the alpha current gain. The injection current lags behind the supply current of the transistor by

$$t_{d\gamma} = (\text{approx.})\ r_{e,j} C_{e,b} \qquad (12\text{-}111)$$

The complex nature of Y_{11b} also suggests that there is a phase shift between the current through and the voltage across the emitter junction. When a current pulse is applied to the emitter, the emitter voltage does not rise instantly. Rather it increases during the time it takes C_e to charge through the emitter resistance and a part of the base spreading (bulk) resistance r_b', which is the resistance between the emitter junction and the base terminal. As a result, a phase shift appears not only between the emitter voltage and the total emitter current, but also between this current and the injection current, because the latter depends on the emitter voltage.

As important are the processes taking place in the base. Let us assume, for example, that the processes in the emitter junction are free from time lag so that, when a current pulse is applied to the emitter, the injection current jumps from zero to I_{ep}.

The injected holes take some time to reach the collector junction because their diffusion time in the base is finite. During the interval $t_{d\alpha}$ when the front of the injected holes moves through the base towards the collector, the collector current does not change at all, but begins to rise as the injected holes diffusing in the base reach the collector junction. Since the thermal velocity distribution of holes is statistical, the time during which the holes reach the collector is not the same. So I_c takes a certain time t_r to approach a steady-state value.

The holes moving in the base also introduce a time lag when the transistor operates, for example, as an r.f. sinusoidal amplifier. If the time during which the injected carriers move from the emitter to the collector is comparable with the period of the signal to be amplified, variations in the concentration of the holes injected by the emitter into the base will not follow a curve drooping monotonically from emitter to collector. Let us assume, for example, that at a certain instant the polarity of the alternating voltage is such that the height of the potential barrier between the emitter and base decreases. If, during the half-cycle, the holes have time to travel only a part of the distance between the emitter and collector, the curve representing the concentration of holes in the base will reach

its maximum somewhere in the middle of the base, because, at that instant, the barrier between the emitter and base will increase and the number of injected holes will decrease considerably. As a result, the forward diffusion of holes in the base will be accompanied by a reverse hole diffusion. The collector current will decrease and, as a consequence, so will the alpha current gain. The time lag introduced in the base is also responsible for the phase shift between I_{ep} and I_{cp}, so that α becomes complex in nature.

Some time lag is likewise introduced by the processes in the base occurring when the current pulse applied to the emitter ceases. The minority carriers previously injected into the base leave the base when I_{ep} falls to zero, but they do so during a certain time t_{dis}, rather than instantaneously. During this time, the minority carriers dissipate from the base through the two junctions and also recombine in the base.

Thus, from the foregoing we may conclude that when the transistor operates with r.f. and pulse signals, it is its current gain that changes most. The current gain, α = (approx.) h_{21b} in the common-base configuration and β = (approx.) h_{21e} in the common-emitter arrangement, becomes complex, and the change in the absolute values, $[h_{21b}]$ and $[h_{21e}]$, is accompanied by that in the phase angle between the input and output currents.

The performance of transistors in r.f. and pulse operation is expressed in terms of a number of parameters. One set of parameters are the time intervals characterizing the waveform of the current pulse in the collector circuit when a square switching pulse is applied to the transistor input. These will be discussed later in the analysis of a transistor operating as a switch.

Another set of parameters are the frequencies at which the current gains fall to certain values.

High-Frequency Parameters. In the general case, h_{21b} may be defined as a function of time by an equation given in reference [24]

$$h_{21b}(t) = h_{21b0}[1 - \exp(-t/t_r)] \qquad (12\text{-}112)$$

This expression does not take into account the delay time $t_d = t_{dv} + t_{d\alpha}$. If we take into consideration t_d, the function (12-112) holds at $t \geqslant t_d$, whereas at $t < t_d$, $h_{21b}(t) = 0$.

The frequency response described by Eq. (12-112) has the following form

$$\dot{h}_{21b} = h_{21b0} \exp[-ic(f/f_{h21b})]/(1 + if/f_{h21b}) \qquad (12\text{-}113)$$

Here, c is the non-dimensional coefficient depending on t_{dv} and varying from 0.25 when $t_{dv} \gg t_r$ to 0.65 when t_{dv} = (approx.) t_r.

The amplitude-frequency and phase-frequency characteristics corresponding to Eq. (12-112) have the form

$$|\dot{h}_{21b}| = h_{21b0}/\sqrt{1+(f/f_{h21b})^2} \qquad (12\text{-}114)$$

$$\varphi h_{21b} = \arctan(f/f_{h21b}) \qquad (12\text{-}115)$$

In Equations (12-113) through (12-115), h_{21b0} is the a.f. current gain.

The quantity f_{h21b} is called the *alpha cut-off frequency* (for the common-base configuration). As follows from Eq. (12-114), f_{h21b} is

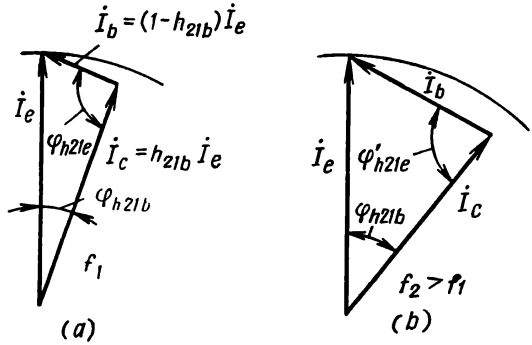


Fig. 12-16. Transistor-current vector diagrams
(*a*) high-frequency; (*b*) higher-frequency

the frequency at which the absolute value of h_{21b} decreases to $1/\sqrt{2}$ of its low-frequency value.

In the common-emitter configuration, the frequency characteristics are similar to those defined by Eqs. (12-113) through (12-115):

$$\dot{h}_{21e} = h_{21e0}\exp[-i(f/f_{h21e})]/(1+if/f_{h21e}) \qquad (12\text{-}116)$$

$$|\dot{h}_{21e}| = h_{21e0}/\sqrt{1+(f/f_{h21e})^2} \qquad (12\text{-}117)$$

$$\varphi_{h21e} = \arctan(f/f_{h21e}) \qquad (12\text{-}118)$$

Here, f_{h21e} is the beta cut-off frequency (for the common-emitter configuration), that is, the frequency at which the absolute value of h_{21e} falls to $1/\sqrt{2}$ of its low-frequency value.

The frequency f_{h21e} is much smaller than f_{h21b}; they are related as follows

$$f_{h21e} = (1-h_{21b0})f_{h21b} = f_{h21b}/(1+h_{21e0}) \qquad (12\text{-}119)$$

As is seen from the above relation, $|\dot{h}_{21e}|$ decreases with frequency faster than $|\dot{h}_{21b}|$, and the phase shift φ_{h21e} between $\dot{I}_b$ and $\dot{I}_e$ is much greater than the phase shift between $\dot{I}_e$ and $\dot{I}_c$. This is illustrated by the current vector diagram shown in Fig. 12-16. As the

frequency increases, the angle φ_{h21b} rises, $|\dot{I}_c|$ falls and, as a consequences, so does $|h_{21b}|$. But $|\dot{I}_b|$ rises faster, so $|h_{21e}|$ falls at the same rate.

The quantities $|h_{21b}|$, $|h_{21e}|$ and φ_{h21b} as functions of frequency are given in Fig. 12-17.

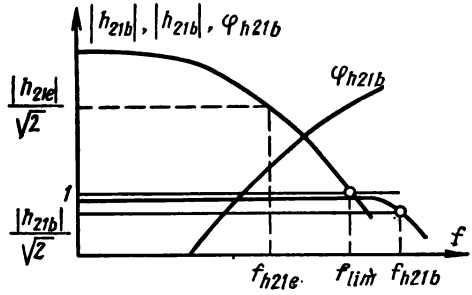


Fig. 12-17. Frequency dependence of h_{21b}, h_{21e} and φ_{h21b}, determining the high-frequency parameters of the transistor (the quantities laid off as ordinates are plotted on a logarithmic scale)

At frequencies several times f_{h21e}, the unity in the denominator of Eq. (12-117) may be neglected, so this expression takes the following form

$$f = h_{21e0} f_{h21e} / |\dot{h}_{21e}| \qquad (12\text{-}120)$$

The frequency $f = f_{lim}$ at which the absolute value of the beta current gain $|h_{21e}|$ decreases to unity is called the *limit frequency* (for the common-emitter configuration)

$$f_{lim} = h_{21e0} f_{h21e} \qquad (12\text{-}121)$$

It is easier to measure f_{lim} than the cut-off frequency, so data sheets usually specify f_{lim}. eferring to Eq. (12-121) and Fig. 12-17, the difference between f_{lim} and f_{h21e} is rather significant.

Another parameter of importance to transistor operation is the *maximum frequency of oscillation*, f_{max}. It characterizes transistor operation as an oscillator. In this case, the power gain

$$K_P = P_{out}/P_{in} \qquad (12\text{-}122)$$

is of importance because when the transistor operates as an oscillator, some of the output power is returned to the input.

In general, if we regard the transistor as an active two-port, K_P may be defined as

$$K_P = v_2 i_2 / v_1 i_1 \qquad (12\text{-}123)$$

Using the equations for the h-parameters and recalling that the output signal voltage is produced across the load resistor R_L connected

in the output circuit ($v_2 = -i_2 R_L$), we may re-write Eq. (12-123) as follows

$$K_P = (\text{approx.})\,(h_{21}^2/h_{11})\,R_L \qquad (12\text{-}124)$$

As is seen, an increase in the signal frequency and a respective decrease in h_{21b} (or h_{21e}) will bring about a decrease in the power gain K_P. When $K_P = 1$, the transistor can no longer be used as an oscillator. This is why the frequency at which $K_P = 1$ is called the *maximum frequency of oscillation.*

The fall in K_P with rising frequency cannot be attributed solely to the decrease in h_{21b} (or h_{21e}). There is also a marked effect of the collector-junction barrier capacitance which is now connected in parallel with the load resistor R_L through a part of the base spreading resistance. At high frequencies, the resistance of this circuit, $r_b' + 1/j\omega C_{c,b}$, decreases and a considerable part of the output current flows through it. Because of this, $f_{\max}$ markedly depends on r_b' and $C_{c,b}$.

As follows from the foregoing, the high-frequency parameters are associated with a particular reduction in the value of the current gain. In turn, the decrease in the current gain with increasing frequency is traceable to the time lag introduced by a number of physical processes in the transistor. The most important among them is the average time for hole diffusion in the base, t_D, provided that $w \ll L_p$ [see Eq. (9-121)]. Accordingly, in the literature of the subject, the high-frequency parameters are sometimes expressed in terms of t_D:

$$f_{h21b} = (\text{approx.})\,1/t_D = (\text{approx.})\,2D/w^2 \qquad (12\text{-}125)$$

$$f_{h21e} = (\text{approx.})\,1/(1 + h_{21e0})\,t_D = (\text{approx.})\,2D/w^2\,(1 + h_{21e0}) \qquad (12\text{-}126)$$

$$f_{lim} = (\text{approx.})\,f_{h21b0}/t_D = (\text{approx.})\,2Df_{h21b0}/w^2 \qquad (12\text{-}127)$$

Large-Signal Operation of the Transistor. This mode of transistor operation is widely used in logical circuits and digital computers. The simplest case is the operation of the transistor switch (Fig. 12-18*a*).

The operation of the transistor switch is illustrated by Fig. 12-18*b* where the load line is constructed on the output characteristics. The manner in which the load line is constructed has been discussed in Sec. 12-6.

If a positive voltage is applied to the input, that is, between the base and emitter, and a negative voltage to the collector with respect to the emitter, the transistor will operate in the cut-off region, and the switch is said to be off. The Q-point is positioned at the intersection of the load line with the $-I_b = I_{cbo}$ curve. The collector current is small, and V_{ce} approaches E_c.

When the input voltage is suddenly reversed in polarity, the emitter junction is forward-biased, and the base current goes posi-

tive. If I_b determined by E_b and R_b is sufficiently high, the Q-point may move along the load line to point B. Thus, the transistor is at first in the active region, then moves into the saturation region. The position of the Q-point at B corresponds to a high collector current and a low V_{ce} (the switch is said to be on).

As is seen, by reversing the polarity of the input voltage or current, we may turn the transistor on (I_c is high) and off ($I_c \to 0$), and utilize it as a static switch in the load curcuit. As a rule, the

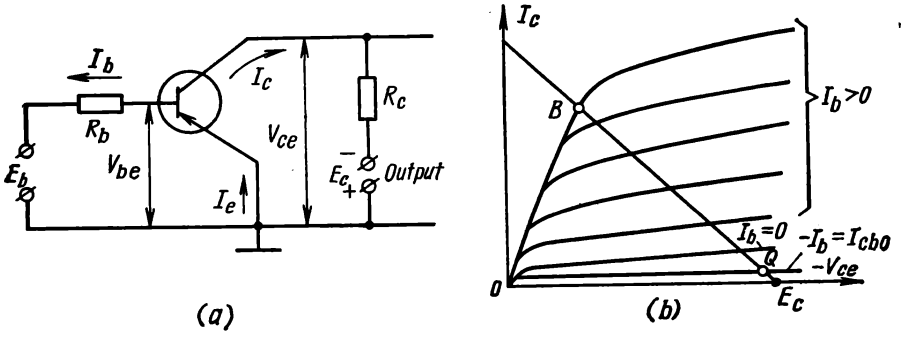


Fig. 12-18. (*a*) Transistor operating as a switch; (*b*) position of the Q-point on the family of output characteristics

transistor is switched by applying an abruptly changing voltage or current, so that the reversal of polarity will occupy a shorter time than the physical processes taking place in the device. Now, let us discuss this matter in more detail.

Figure 12-19*a* shows the waveforms of base and collector currents as the transistor is switched from the off- to the on-state and back. Figure 12-19*b* illustrates the distribution of minority curriers in the base at different moments.

During the time $0—t_1$, the transistor is in the cut-off region, and $I_c = I_{cbo}$. In this case, the distribution of minority carriers is represented by curve *1*. At the instant t_1, I_b suddenly reverses its polarity, but the collector current begins to rise noticeably only after the time interval $t_d = t_{d\gamma} + t_{d\alpha}$. As already noted, t_r depends on the statistical spread in velocity of the holes diffusing in the base.

As C_e charges, the emitter junction is driven into conduction still more, the number density of holes in the base near the emitter junction increases (curves *2*, *3* and *4* in Fig. 12-19*b*) and so does the collector current. During the time t_r, a hole charge is built up in the base, and the hole concentration at the collector junction reaches the steady-state value, p_{n0} (curve *4*).

The increase in I_c is, however, limited by the load resistor R_L in the collector circuit. So the number of holes leaving the collector for the external circuit is smaller than the number of holes entering

the collector from the base. As a result, the collector accumulates holes until they balance the negative space charge at the collector junction, the potential barrier becomes lower, the collector junction is rendered conducting, and the transistor moves into the saturation region. As this takes place, the number density of holes in the base near the collector junction exceeds the steady-state value p_{n0} (curve *5* in Fig. 12-19*b*), and the collector circuit carries a saturation current whose value depends on E_c and R_L, because the resistance of the

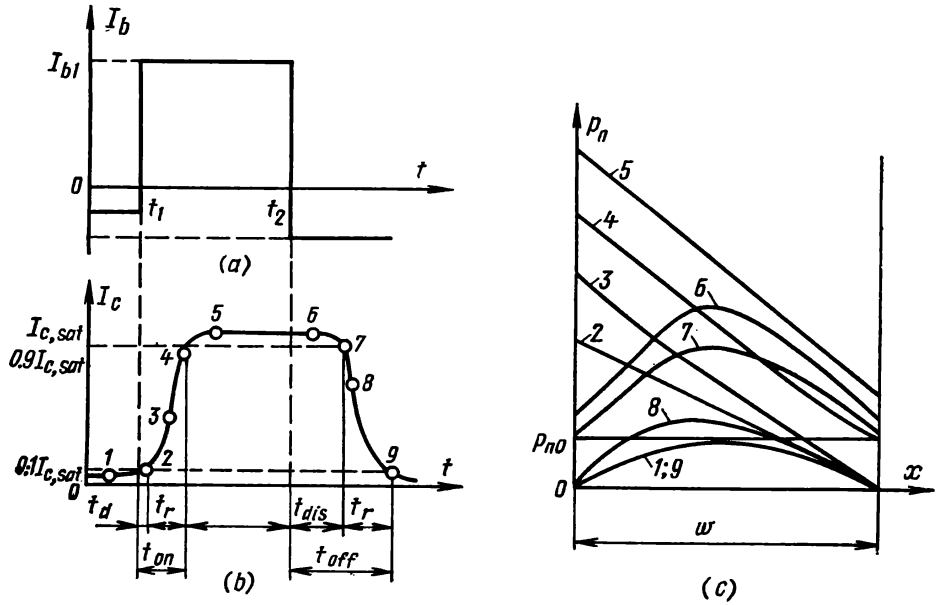


Fig. 12-19. (*a*) Base-current waveform; (*b*) distribution of minority carriers in the transistor base; (*c*) collector current waveforms

collector junction in the on-state is much smaller than R_L: $I_{c,sat} =$ $=$ (approx.) E_c/R_L.

In the saturated region, the collector voltage decreases to $V_{ce,\ sat}$. Thus, in the saturated region, $I_{c,\ sat}$ remains unchanged, although the base continues to build up the hole charge.

At the instant t_2, I_b abruptly reverses its polarity, and the holes stored in the base begin to dissipate. The hole concentration in the base at the emitter and collector junctions, however, cannot decrease to zero in the same abrupt manner. This would produce infinite currents in the emitter and collector circuits, because a considerable charge would have to be transferred from the base into the emitter and collector during an infinitesimal interval of time. Actually, the currents accross the junctions are limited by R_b and R_L, so the number density of holes at the emitter and collector junctions goes down gradually (curve *6*) during the time t_{dis}. Until the hole con-

centration at the junctions reaches a steady-state value, the transistor continues to operate in the saturated region, and I_c remains practically unchanged (the hole concentration gradient at the collector junction for curves *6* and *7* remains practically constant).

After the concentration p_n has reached the steady-state value at the emitter or collector junction or both, the junction or junctions become reverse-biased. The transistor moves into the cut-off region, and tne hole number density at the junctions approaches zero (curve *8*). The concentration gradient at the junctions decreases, the rate of hole dissipation rises, and I_c rapidly falls during the time t_f.

As follows from the foregoing, the operation of the transistor as a switch is affected by the same physical quantities that introduce a time lag in r.f. operation, namely the emitter and collector capacitances C_e and C_c, the hole diffusion constant D_p for the base, the hole lifetime τ_p for the base, etc.

Pulse Parameters. The operation of the transistor as a switch is expressed in terms of the time intervals shown in Fig. 12-19*a*.

The *delay time*, t_d, is the interval between the instant the transistor is turned on and the instant when I_e reaches 0.1 $I_{c,\ sat}$.

The *rise time*, t_r, is the interval during which I_c rises from 0.1 $I_{c,\ sat}$ to 0.9 $I_{c,\ sat}$.

The *storage time*, t_s, is the interval between the instant I_b reverses its polarity and the instant when the collector voltage reaches a specified value.

The *fall time*, t_f, is the interval during which I_c falls from 0.9 $I_{c,sat}$ to 0.1 $I_{c,sat}$.

The pulse parameters also include the *turn-on time*, $t_{on} = t_d + t_r$, and the *turn-off time*, $t_{off} = t_s + t_f$.

12-8. EFFECT OF THE OPERATION MODE AND TEMPERATURE ON TRANSISTOR CHARACTERISTICS AND PARAMETERS

The Effect of the Mode of Operation. When discussing the physical processes in and parameters of the transistor (Sec. 12-2) we derived the expressions relating these parameters to the currents and voltages in the transistor. In Sec. 12-6, we have examined the relations between the physical and *h*-parameters of the transistor. Now, using the above relations, we can evaluate the effect that the operation mode can have on the physical and *h*-parameters of the transistor. Averaged plots of the main parameters as functions of V_{cb} and I_e are shown in Figs. 12-20 and 12-21.

The values of these parameters relative to their values at $V_{cb} =$ $= 5$ V and 2 V, and $I_e = 1$ mA are laid off as ordinates. The curves may de described by the respective equations given in Sec. 12-2.

The Effect of Temperature. Variations in ambient temperature affect the static characteristics and parameters of the transistor rather significantly. It is natural because variations in temperature bring about variations in the concentration of free carriers, their mobility, diffusion constants, recombination probability, etc. This is described by appropriate expressions given in Chapter 9. The effect of temperature on the physical parameters is defined by the equations derived in Sec. 12-2. Using these equations, we may also explain the effect of temperature variations on the h-parameters or static characteristics of the transistor.

In the common-base configuration, the static input ($V_{cb} = 0$) and output ($I_e = 0$) characteristics are similar to the forward and

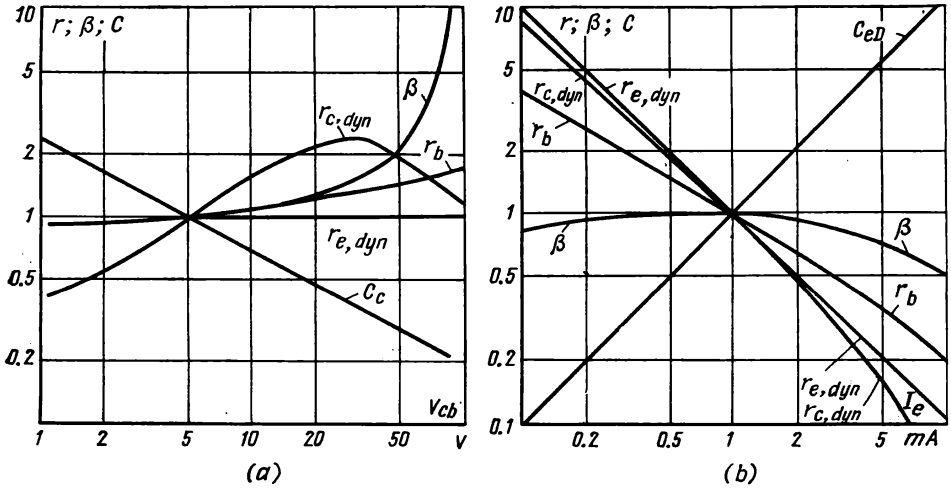


Fig. 12-20. (a) Physical parameters of the transistor as functions of V_{cb} and (b) as functions of emitter current

reverse slopes of the static characteristic of the diode. Accordingly, the reasoning set forth in Sec. 11-3 holds for the static characteristics of the transistor. Above all, we may conclude that instability in the characteristic is mostly caused by the reverse current, I_r, or (for the transistor) the collector leakage (or cutoff) current, I_{cbo}. The collector current as a function of temperature is described by an expression similar to Eq. (11-6). In germanium transistors, the main contributor to the I_{cbo} is the thermal current whose value doubles every 8 to 10°C of temperature rise. In silicon transistors, the main component of I_{cbo} is the generation current I_g. This current doubles about every 10°C of temperature rise.

According to Eq. (12-33), the dynamic resistance of the emitter junction is also directly proportional to T. Rising temperature

affects the concentration of carriers, so the base spreading (bulk) resistance r_b changes. The processes in the base and the changing carrier concentration at the collector junction bring about variations in the alpha and beta current gains and the dynamic resistance of

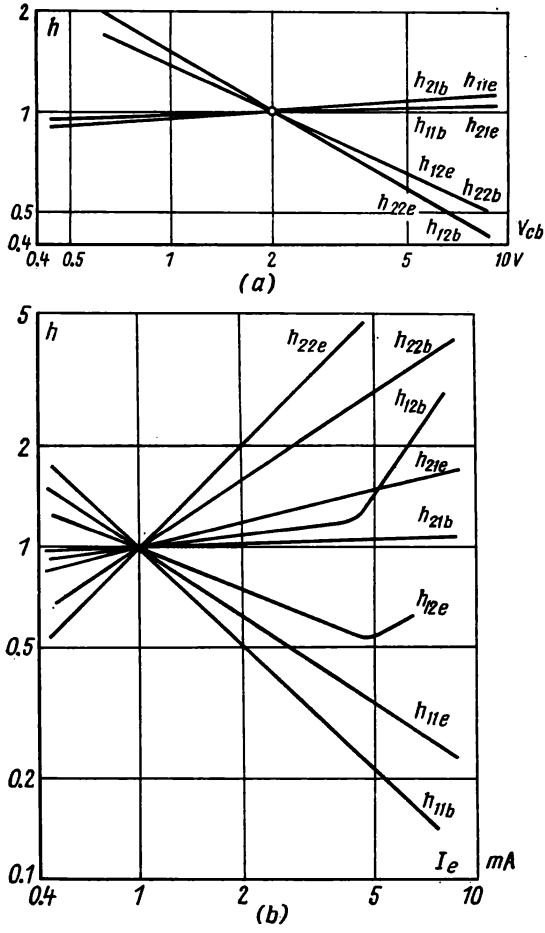


Fig. 12-21. (*a*) *H*-parameters as functions of V_{cb} and (*b*) as functions of emitter current

the collector junction $r_{c,dyn}$. Variations in these parameters and the related *h*-parameters are illustrated in Figs. 12-22*a* and 12-22*b*, respectively.

The effect of temperature variations on the output characteristics of the transistor connected in a common-emitter circuit is illustrated in Fig. 12-23.

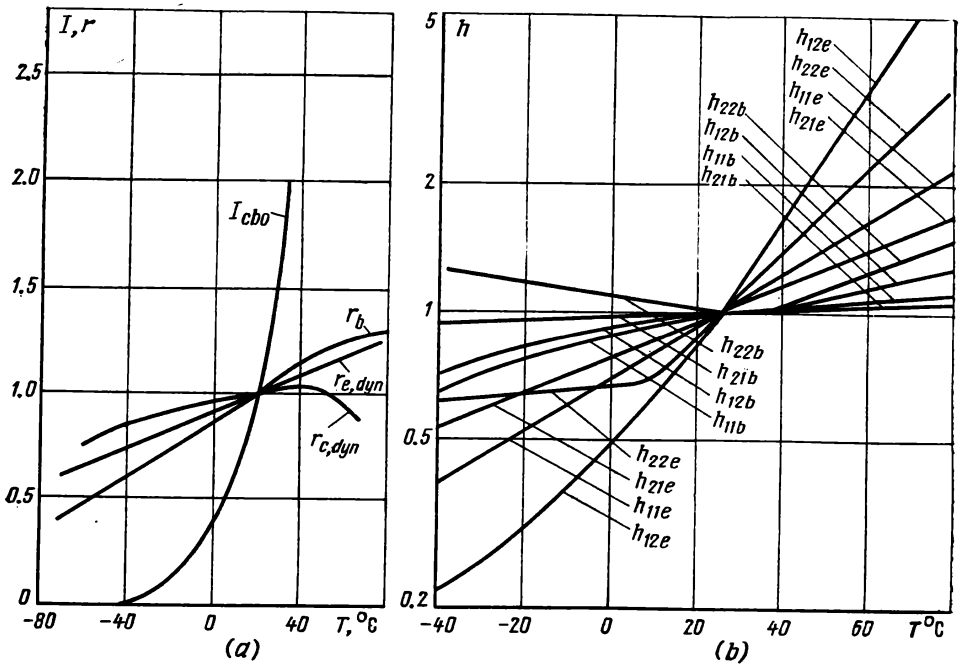


Fig. 12-22. (*a*) Physical parameters as functions of temperature in a steady-state; (*b*) h-parameters as functions of temperature in a steady-state

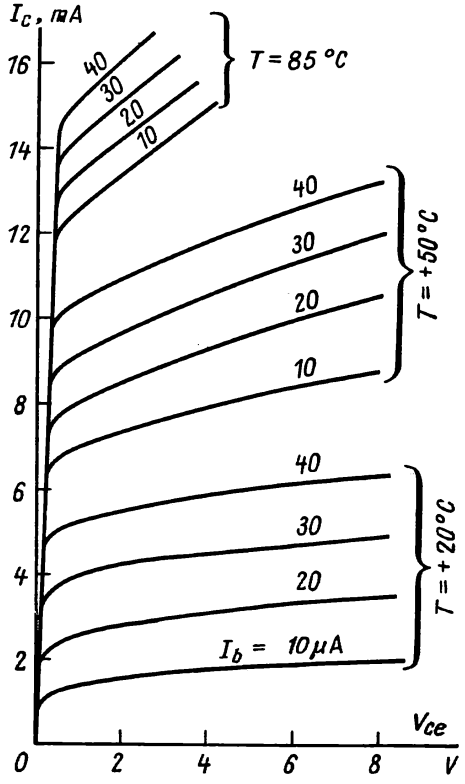


Fig. 12-23. Effect of temperature variations on the output characteristics of a common-emitter transistor

12-9. CONSTRUCTION AND TYPES OF BIPOLAR TRANSISTORS

Bipolar transistors are many and diverse. They differ in purpose, main parameters, material, methods of manufacture, and some other aspects.

Most transistors are made of silicon and germanium, although gallium arsenide is also used.

Transistors are manufactured of two main types, namely *P-N-P* and *N-P-N*.

In accordance with their power rating, transistors are customarily divided into three groups, namely low-power transistors with the collector power dissipation $P_c \leqslant 0.3$ W, medium-power devices with $0.3 \text{ W} < P_c \leqslant 5$ W, and high-power transistors with $P_c > 5$ W. In contrast to low-power transistors, the last two types can handle heavy currents and dissipate considerable thermal power.

The three groups may be subdivided in accordance with the frequency range or the method of manufacture.

In accordance with the frequency range, transistors are classified into four types: low-frequency transistors with the alpha cut-off frequency $f_{h21b} \leqslant 3$ MHz, medium-frequency transistors ($3 \text{ MHz} < f_{h21b} < 30$ MHz), high-frequency transistors ($30 \text{ MHz} < f_{h21b} \leqslant 120$ MHz), and microwave transistors ($f_{\max} > 120$ MHz).

Let us discuss the distinctions of the various transistors and the main methods by which the necessary parameters are obtained.

Low-Frequency Low-Power Transistors. As a rule, low-power transistors are manufactured by fusing pellets of alloying material into a semiconductor (germanium or silicon) wafer serving as a base. These are *alloy* transistors.

The alloying pellets used to form the emitter and collector junctions in germanium transistors are prepared from indium with an addition of gallium (for *P-N-P* types) and an alloy of antimony and zinc (for *N-P-N* types). *P-N-P* silicon transistors use aluminium, and *N-P-N* types an alloy of phosphorus and zinc.

The pellets are fused into the semiconductor wafer and then fired at 500-600°C in an atmosphere of hydrogen. As a result, the structure shown in Fig. 12-24*a* is formed. After that, the structure is attached to supports (serving as leads) pressed into the stem (bottom) of the device and is enclosed in an air-tight case (Fig. 12-24*b*). The drawback of this method is that the thickness of the base is difficult to control.

Low-frequency transistors have considerable junction capacitances ($C_{c,b} \approx 10$ to 200 pF) and relatively high I_{cbo} reaching 10 to 15 μA in some types.

High-Frequency Low-Power Transistors. As follows from the analysis of the transistor operation with high-frequency signals

(Sec. 12-7), the cut-off frequency depends on a number of physical parameters. In order to raise this frequency and, as a consequence, extend the high-frequency end of the spectrum, the junction capacitances should be reduced, which is achieved by reducing the junction areas, the bulk resistances of the emitter, base and collector, and the thickness of the base, or by some other measures accelerating the motion of carriers in the base.

In the alloy technique, the base region cannot be reduced significantly because, in the course of alloying, the depth of the emitter and collector junctions cannot be controlled with the desired accu-

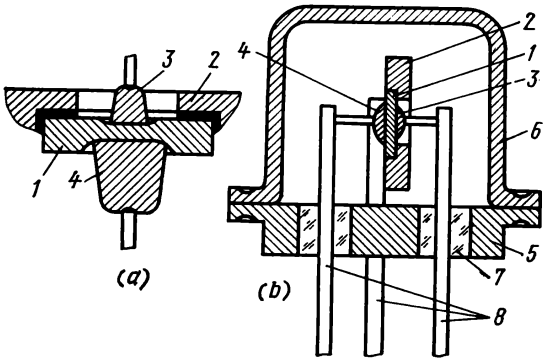


Fig. 12-24. Construction of a low-frequency alloy transistor

(*a*) semiconductor crystal with fused-in pellets; (*b*) transistor enclosed in a case; *1*—semiconductor wafer; *2*—support; *3*—emitter; *4*—collector; *5*—stem; *6*—case; *7*—insulator; *8*—leads

racy. It is also impossible to reduce considerably the emitter and collector junction areas.

Because of this, high-frequency transistors are manufactured mainly by the alloy-diffusion and planar methods.

As its name implies, in the alloy-diffusion technique, the pellets of impurity material are fused into the semiconductor wafer serving as a collector and then the impurity atoms diffuse into it.

For example, the pellets for germanium alloy-diffused transistors are dots of germanium doped with both donor (antimony) and acceptor (indium) impurities. When the pellet is fused into the *P*-type germanium wafer, the atoms of both impurities diffuse into the wafer, but the diffusion length and solubility for antimony and indium are different. As a result of fusion, the antimony atoms penetrate into the wafer more deeply, thereby forming the *N*-region of the base, whose impurity concentration decreases with increasing distance from the surface.

At the same time, the indium atoms penetrate into the germanium to a smaller depth, but, since the solubility of indium in germanium is much higher, the concentration of acceptor impurities near the

fusion area is considerably higher than that of donor impurities. As a result, a highly-alloyed P-type emitter region is formed near that area.

The concentration of the acceptor and donor impurities in such a transistor is illustrated in Fig. 12-25*a*; Fig. 12-25*b* shows the curve representing the difference in concentration between the acceptor and donor impurities, $N_a - N_d$, and Fig. 12-25*c* gives the energy-band diagram of the transistor.

One of the principal features of this method is that the base is nonuniformly alloyed: the donor impurity concentration which is sufficiently high at the emitter junction gradually decreases towards the collector. As a result, the concentration of majority carriers (electrons) in the base is nonuniform, being considerably higher at the emitter junction. Owing to the arising concentration gradient, the electrons in the base diffuse towards the collector junction, leaving behind at the emitter junction the positive charges of ionized donor atoms. This sets up an electric field $\mathscr{E}_b$ whose strength vector is directed from the emitter to the collector (Fig. 12-25*b*). Under the action of this field, the motion of minority carriers (holes) injected from the emitter into the base is caused not only by diffusion as in a conventional alloy transistor with a uniformly alloyed base, but also by drift. These are *drift transistors.*

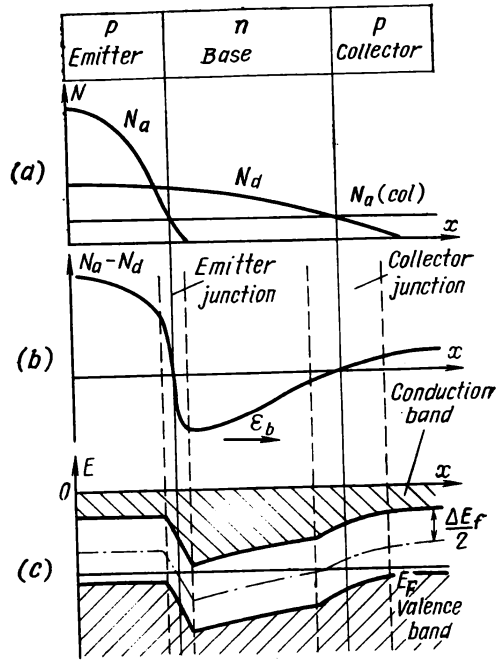


Fig. 12-25. Drift transistor

(*a*), (*b*) impurity concentration; (c) energy-band diagram

A word of caution is in order. Transistors having a uniformly alloyed base and minority carriers transported only by diffusion are called *diffusion transistors.* The diffusion type of transistors ought not to be confused with *diffused* transistors, because the word 'diffused' refers to the method of production and not to the mechanism of carrier transport.

Drift transistors can also be manufactured by the planar technology. The construction of a planar transistor is shown in Fig. 12-1*b*. In this method widely used in the production of integrated circuits, the base and emitter regions are formed in a semiconductor wafer by local diffusion. At first, a wafer of N-type silicon is oxidized,

then on its surface is manufactured a thin film of silica (SiO_2) which has dielectric properties and stands up well to attack by the atmosphere and many chemicals.

Next, photoetching is used to remove the film of SiO_2 from the place where the base region (P-type silicon) is planned to be. By this method, the film may be removed from very small areas (about several tenths of a square micrometer) with high precision. Then, through this portion (called the window), an acceptor impurity is diffused in special ovens, and the base region is formed. After tnat, a new oxide layer is grown over the new silicon surface, a new smaller area is cleared, and a donor impurity is diffused through the new window to form the emitter region.

As a result, the impurities are arranged in a manner similar to that shown in Fig. 12-25*b*.

The above technique is efficient not only in making a nonuniform concentration of impurities in the base. It also helps to improve some other transistor parameters and extend the high-frequency end of the operating frequency range.

In these production methods, the diffusion is more precisely controlled and the base width can be reduced to a split micrometre, so that it is 1/10 or 1/15 of the value in alloy transistors. As a result, the diffusion time for holes in the base t_D defined by Eq. (9-121), the rise time t_r, and, as a consequence, the cut-off and limiting frequencies defined by Eqs. (12-125) through (12-127) are significantly decreased.

Because of the decrease in t_r, the current gain h_{21e} may reach several hundreds.

Naturally, the field present in the base also serves to improve the high-frequency parameters of the drift transistor. As a rule, the drift time for minority carriers in the base, t_d, is half to one third of t_D.

Due to the high impurity concentration in the emitter and the base near the emitter junction in drift transistors, the emitter junction is very narrow, so the barrier capacitance $C_{e,\,b}$, it would appear, must be high. The above manufacturing techniques, however, make it possible to reduce significantly the junction areas, especially that of the emitter, so $C_{e,\,b}$ is made small.

The collector barrier capacitance, $C_{c,\,b}$, can also be considerably lowered (to about 1/10 as compared with alloy transistors, because the collector junction is rather wide owing to the low impurity concentration in the collector and the base near the collector junction.

Because of this, drift transistors can be used at frequencies up to several gigahertz.

Since the emitter junction is rather narrow, its peak-reverse voltages are small, being 1 to 2 V. At high V_{eb}, the Zener breakdown

occurs, which, at low currents, cannot destroy the junction. The collector junction is rather wide and has a relatively high dynamic resistance. Because of this, the maximum safe power dissipation may be raised to 100-150 MW.

High-frequency low-power transistors use cases of about the same design as low-frequency devices. Mostly, they are also enclosed in

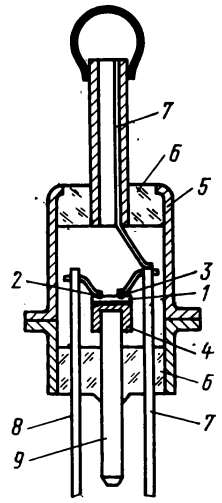


Fig. 12-26. High-frequency alloy-diffused transistor
1—collector; *2*—base; *3*—emitter; *4*—support; *5*—case; *6*—insulator; *7*—emitter lead; *8*—base lead; *9*—collector lead

air-tight metal-glass cases. The only difference is the manner in which the transistor is attached to the support. In high-frequency transistors, the semiconductor wafer usually serves as a collector, and this is fastened to the support.

Sometimes, transistors are enclosed in cases (Fig. 12-26) suitable for connection in coaxial circuits.

High-Power Transistors. High-power transistors are designed and produced in such a way that they can handle high currents, operate at increased voltages, and dissipate a considerable amount of thermal energy.

Most high-power transistors are manufactured by the alloy process. High-frequency high-power transistors are fabricated by the planar technique.

The distinctions of high-power alloy transistors are as follows. The emitter is made by fusing indium with an addition of gallium, which raises the injection ratio and emitter currents. In order to reduce the base resistance and increase the emitter area for higher I_e, use is made of strip emitter junctions (Fig. 12-27*a*). For better heat abstraction, the collector is soldered to the case base which is sometimes equipped with a special heat sink (see Fig. 9-1*c*).

Using the planar production method, it is possible to make high-power multiemitter transistors. In these transistors, the surface of

the base region carries many emitter regions made in the form of strips or rings separated from one another. They are connected by

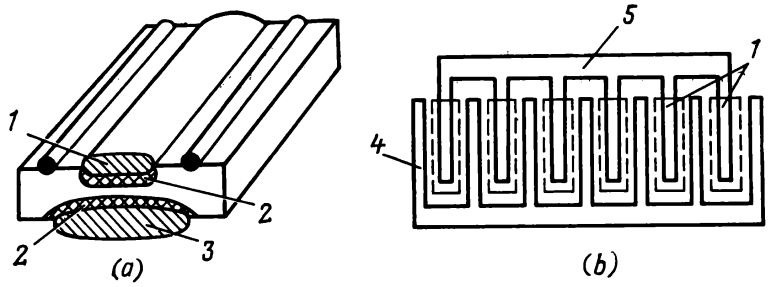


Fig. 12-27. Electrodes of high-power transistors

(*a*) alloy transistor; (*b*) planar transistor; *1*—emitter; *2*—*P-N* junction; *3*—collector; *4*—base lead; *5*—emitter lead

a metal-film strip vacuum-deposited on the crystal surface (Fig. 12-27*b*).

Chapter Thirteen

FIELD-EFFECT TRANSISTORS

13-1. GENERAL

Definition. *A field-effect transistor* (*FET*) is a semiconductor device with the electrodes called the *source*, *gate* and *drain*. The current through the semiconductor region called the *channel* is controlled by the electric field established by a voltage applied between the gate and the source.

The channel is the semiconductor region between the source and the drain in which the charge-carrier flow is controlled by varying its cross-sectional area. The source is the electrode supplying charge carriers into the channel. The drain is the electrode through which carriers leave the channel. And, finally, the gate is the electrode or electrodes to which a bias is applied for the purpose of modulating the conductivity of the channel.

In the literature of the subject, field-effect transistors are sometimes called *unipolar* transistors because the current flow in them depends on the movement of majority carriers only, in contrast to bipolar transistors which use both negative and positive charge carriers.

Classification. There are two main types of FETs, namely *junction field-effect transistors*, (JFET)s, and *insulated-gate field-effect transistors*, (IGFET). Of late, a good deal of prominence has been given to metal-oxide-semiconductor insulated-gate field-effect transistors (MOSFET) which use a metal gate separated from the channel material by a thin layer of silicon dioxide (SiO_2). In turn, MOSFETs fall into two types, namely a *depletion-mode* MOSFET and an *enhancement-mode* MOSFET.

Also, all field-effect transistors are classified according to the channel conduction into *P*- and *N*-channel devices.

Design. Figure 13-1 shows the construction of three types of FET. In the JFET, the channel is formed by a part of the *N*-type semiconductor crystal having a smaller cross-sectional area. The *N*-type semiconductor has two *P*-regions which border on the *N*-type crystal, thereby constituting a *P-N* transition region.

In the depletion-mode MOSFET (Fig. 13-1*b*), the *P*-channel and the drain and source (P^+-type) regions are formed during the manufacture of the device. In the enhancement-mode MOSFET (Fig. 13-1*c*), the *P*-channel is produced after the transistor has been fabricated. For this purpose, a voltage is applied to the gate, which develops an electric field forming a thin inverted layer with hole conduction near the surface of the *P*-type semiconductor between the source and drain.

The JFET may also be manufactured of *P*-type semiconductor (*P*-channel), and the IGFET may be made with the N^+-type source and drain regions produced in a crystal having hole conduction.

Referring to Fig. 13-1*b* and *c*, in the IGFET, the gate is insulated from the crystal by a dielectric layer. This is why they are called insulated-gate field-effect transistors.

The MOSFET has four electrodes, the fourth being the *substrate* of the transistor.

Major Distinctions. As already noted, in contrast to bipolar transistors, the current in field-effect transistors is only due to the

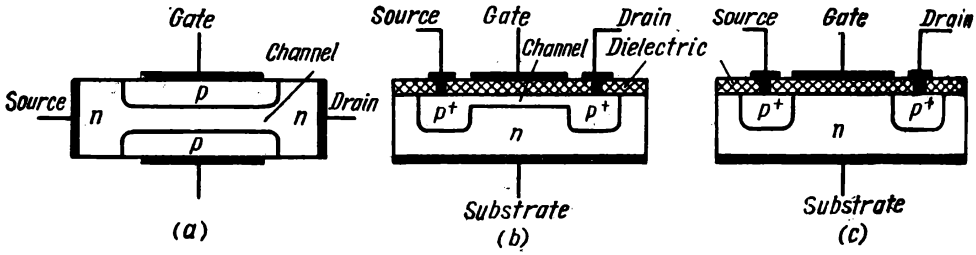


Fig. 13-1. Construction of field-effect transistors
(*a*) JFET; (*b*) depletion-mode MOSFET; (c) enhancement-mode MOSFET

drift of the majority carriers. Accordingly, the frequency properties and behaviour of field-effect transistors in pulse work are affected by physical parameters and processes other than in bipolar transistors.

Another distinction is that the current in the FET is controlled by an electric field. This field is developed by a reverse voltage applied to the control *P-N* junction in the JFET or to the gate in the IGFET. In either case, the control and gate currents are rather small, and the input dynamic (a.c.) resistance of the transistor is high, being from 10^8 to 10^{10} ohms in the JFET, and from 10^{10} to 10^{12} ohms in the IGFET. In this respect, field-effect transistors are similar to vacuum tubes. Because of this, the performance of the FET as an amplifier is customarily described, as with vacuum tubes, in terms of its transconductance expressing the drain (output) current as a function of the gate (input) voltage, rather than in terms of the current gain as in the bipolar transistor.

The diagram symbols for Soviet-made field-effect transistors are set up by USSR State Standard GOST 2.730-73, and the nomenclature and definitions are established by GOST 19095-73.

13-2. JUNCTION FIELD-EFFECT TRANSISTORS

Construction. Connection in a Circuit. The construction of a junction field-effect transistor (JFET) is shown in Fig. 13-1*a*. Like the bipolar transistor, the FET has three circuit configurations (Fig. 13-2),

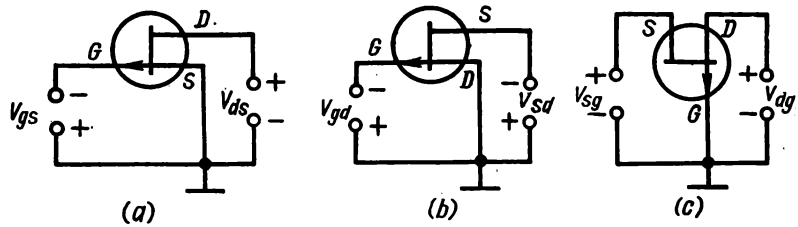


Fig. 13-2. JFET (*N*-channel) circuit configurations
(*a*) common-source; (*b*) common-drain; (*c*) common-gate

namely *common-source*, *common-drain*, and *common-gate*. Most often, use is made of the common-source configuration.

Principle of Operation. The operation of the JFET is schematically shown in Fig. 13-3. Here, the JFET is connected in a common-

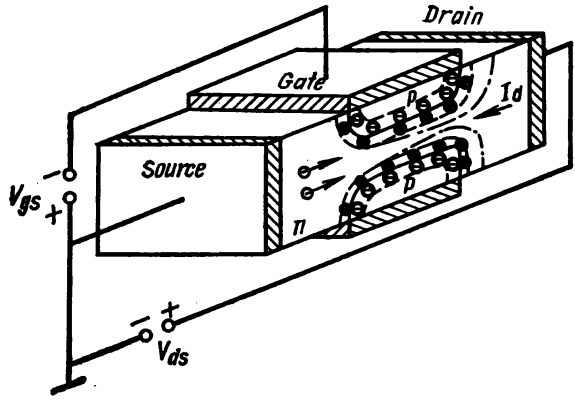


Fig. 13-3. Operating principle of a JFET

source circuit. First, let us assume that the voltages between the transistor electrodes are equal to zero. At the interface between each *P*-region and the *N*-type semiconductor crystal, or the channel, there is a *P-N* junction in which the depletion-region width is determined by Eq. (10-24). The resistance of the channel depends on the cross-

sectional area of the narrow N-type semiconductor region between the upper and lower P-N junctions. When the P-N junction formed by the gate and source is reverse-biased, that is, V_{gs} is negative, the depletion region will broaden according to Eq. (10-37), the channel will decrease in cross-section and, as a consequence, its resistance will rise. So, by varying V_{gs} we may control the channel resistance. According to Eq. (10-37), the effect of variations in V_{gs} increases with rising resistivity of the N-type crystal.

When a positive drain-to-source voltage, V_{ds}, is applied, majority charge carriers—electrons—will drift from the source through the channel to the drain, thereby giving rise to the drain current, I_d. The positive drain-to-source voltage also changes the channel configuration. If we neglect the voltage drop across the bulk resistances of the N-type semiconductor regions between the source and channel, and between the drain and channel, the channel potential at the source will be zero, and at the drain it will be equal to V_{ds}. The voltage across the P-N junction at the source will be $|V_{gs}|$, and that at the drain $|V_{gs}| + V_{ds}$; as a result, the depletion region at the drain will broaden, as shown by the dash-dot line in Fig. 13-3.

Thus, when V_{ds} is positive, there is a flow of drain current, I_d, whose value is determined by the total resistance of the channel having nonuniform cross-sectional area. When V_{ds} is negative, the potential barrier builds up and does not allow the electrons travelling via the channel to move to the P-regions. Now the current through the P-N junctions and, as a consequence, the gate current is solely due to the minority carriers, which, at low values of n_p in the P-regions and of p_n in the N-type crystal, is rather small.

The drain current, I_d, may be controlled by varying V_{gs} or V_{ds}. At a certain negative value of V_{gs}, the depletion regions of the upper and lower P-N junctions may come together. The channel is said to be pinched off at that point, and the transistor is turned off, that is, $I_d = 0$. The value of V_{gs} at which the transistor is cut off is called the *pinch-off voltage*, designated $V_{gs,p\text{-}off}$.

Pinch-off may also occur when V_{ds} is raised. In this case, however, $I_d \neq 0$. This will be discussed later in the section dealing with FET characteristics.

Static (D.C.) Characteristics. Like the bipolar transistor, the FET may be represented by an equivalent two-port. In small-signal operation, such a two-port may be considered to be linear. Because the FET is a voltage-controlled device, so, like vacuum tubes, it may be described by a system of equations using the y-parameters:

$$\begin{aligned} di_1 &= y_{11}\, dv_1 + y_{12}\, dv_2 \\ di_2 &= y_{21}\, dv_1 + y_{22}\, dv_2 \end{aligned} \tag{13-1}$$

For the most widely used common-source circuit configuration, $i_1 = I_g$, $v_1 = V_{gs}$, $i_2 = I_d$ and $v_2 = V_{ds}$.

The y-parameters are the main parameters of the field-effect transistor.

According to Eqs. (13-1), the following relations between the currents and voltages of the FET may be used as its static characteristics:

—input characteristic

$$I_g = f_1/(V_{gs}), \quad V_{ds} = \text{constant} \tag{13-2}$$

—reverse transfer characteristic

$$I_g = f_2(V_{ds}), \quad V_{gs} = \text{constant} \tag{13-3}$$

—forward transfer characteristic

$$I_d = \varphi_1(V_{gs}), \quad V_{ds} = \text{constant} \tag{13-4}$$

—output characteristic

$$I_d = \varphi_2(V_{ds}), \quad V_{gs} = \text{constant} \tag{13-5}$$

As a rule, use is only made of the last two characteristics. Let us discuss them in more detail.

Output (Drain) Characteristics. These are the main characteristics of the field-effect transistor (Fig. 13-4).

Let us examine the $I_d = f(V_{ds})$ curve at $V_{gs} = 0$. At first, when V_{ds} is low, I_d rises almost linearly with increasing V_{ds}. The slight departure of this region from the straight line is accounted for by the fact that the channel becomes narrow at the drain where the junction voltage is $|V_{gs}| + V_{ds}$, so that its resistance increases. This is known as the *ohmic region*. As V_{ds} increases, the curves flatten out and the region is termed the *saturation region*; here, I_d will not increase further with increasing V_{ds}. This occurs because of pinch-off at the drain. When V_{ds} is sufficiently high, the channel shrinks so that it forms a very narrow strip (neck). Under the circumstances, a kind of dynamic equilibrium exists: as V_{ds} and I_d increase the channel width decreases, so the current is reduced. The value of the drain voltage at which this takes place is termed the *saturation voltage*, designated $V_{ds,sat}$.

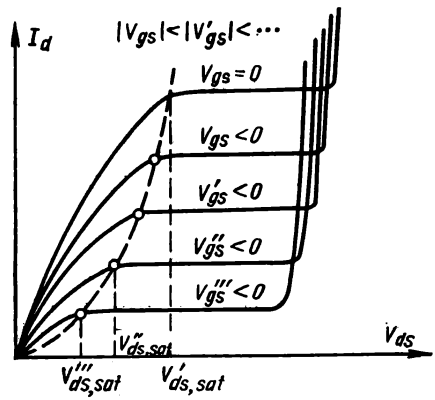


Fig. 13-4. Output (drain) characteristics of a JFET

As is seen, when $V_{ds} = V_{ds,sat}$ and $|V_{gs}| < |V_{gs,p\text{-}off}|$, the transistor behaves as when $V_{gs} = -V_{gs,p\text{-}off}$, but at $V_{ds} < V_{ds,sat}$. The only difference is that in the latter case $R_{ch} \to \infty$ and the tran-

sistor is cut off, whereas in the saturation region the dynamic channel resistance, $r_{ch,\,dyn} = dV_{ds}/dI_d$, tends to infinity. Since V_{gs} and V_{ds} affect the channel width at the drain almost equally, we may write

$$V_{ds,\,sat} = |V_{gs,\,p\text{-}off}| - |V_{gs}| \qquad (13\text{-}6)$$

A certain rise in current in the saturation region can be explained as follows. As V_{ds} increases, the channel neck grows longer (Fig. 13-5). In addition, with rising V_{ds}, the field in the channel neck reaches a value at which the electron mobility remains almost unchanged with increasing field strength. This contributes to the channel resistance which, however, builds up slowly rather than in proportion to increasing V_{ds}. As a result, I_d slightly rises in the saturation region all the same.

When V_{ds} is raised sufficiently high, I_d suddenly rises. This happens because a maximum reverse voltage is now applied to the P-N junction at the drain end of the channel, and causes the junction to break down.

When the gate is held at a negative voltage, the saturation region takes place at lower drain voltages ($V''_{ds,\,sat} < V'_{ds,\,sat}$); I_d is also lower because the cross-sectional area of the channel decreases along its full length; and, finally, the P-N junction breaks down at lower V_{ds}.

The output (drain) characteristics of the JFET resemble the plate characteristics of the pentode. As with the pentode, the operating region is represented by the gradual portions of the characteristics at $V_{ds} > V_{ds,\,sat}$.

The $I_d = \varphi_2(V_{ds})$ relation is described by an equation given in reference [24]

$$I_d = (1/r_{ch.dyn0})\,\{V_{ds} + (2/3)[|V_{gs}|^{3/2} - (|V_{gs}| + V_{ds})^{3/2}]/\sqrt{|V_{gs,p-off}|}\} \qquad (13\text{-}7)$$

Here, $r_{ch,\,dyn0} = \partial V_{ds}/\partial I_d$ is the minimum dynamic channel resistance at $V_{gs} = 0$ and $V_{ds} = 0$.

Substituting Eq. (13-6) into the above relation gives the expression for the saturation current

$$I_{d,sat} = (1/r_{ch,dyn0})\,\{(1/3)\,|V_{gs,p-off}| - |V_{gs}|\,[1 - (2/3)\sqrt{|V_{gs}|/|V_{gs,p-off}|}]\} \qquad (13\text{-}8)$$

Forward Transfer (Drain-Gate) Characteristic. This is shown in Fig. 13-6. The current begins to flow in the JFET at $|V_{gs}| < |V_{gs,\,p\text{-}off}|$. In the saturation region, V_{ds} does not practically affect this characteristic which is quite accurately described by Eq. (13-8) or a simpler quadratic voltage dependence of current given in reference [24]

$$I_{d,sat} = (\text{approx.})\,(1/r_{ch,dyn0})\,[(|V_{gs,p-off}| - |V_{gs}|)^2/3\,|V_{gs,p-off}|] \qquad (13\text{-}9)$$

Input Characteristic. This characteristic of the JFET (Fig. 13-7) is the downward slope of the volt-ampere characteristic of the *P-N* junction. The gate (input) current depends on the drain-to-source

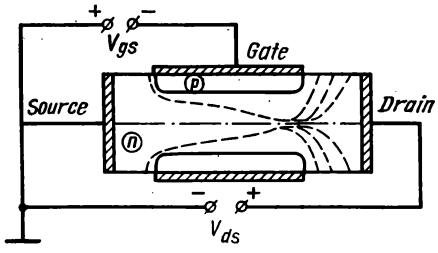


Fig. 13-5. Effect of increasing V_{ds} on the channel shape

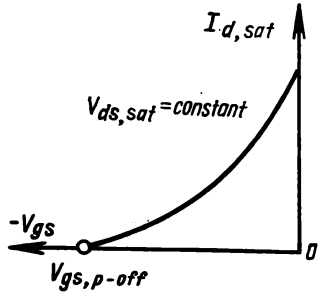


Fig. 13-6. Forward transfer (drain-gate) characteristic of a JFET

voltage. It reaches its maximum when the source and drain are short-circuited; $I_g = I_{g,l}$ (where $I_{g,l}$ is the gate leakage current).

When V_{gs} is negative and the input is open-circuited, the reverse current I_{gso} flows from the drain to source and from the drain to gate, and the reverse current I_{gdo} flows from the gate to drain.

In either case, the reverse current through the junction is slightly affected by $|V_{gs}|$, which, as will be recalled, is typical of the downward slope of the *P-N* junction characteristic. Because of this, the input dynamic resistance of the transistor, $R_{in,\,dyn} = \partial V_{gs}/\partial I_g$, is very high.

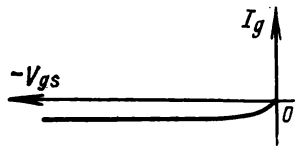


Fig. 13-7. Input characteristic of a JFET

The parameters of the JFET will be discussed at the end of this chapter because they are the same for all types of field-effect transistors.

13-3. INSULATED-GATE FIELD-EFFECT TRANSISTORS

The Depletion-Mode MOSFET. The construction of this transistor is shown in Fig. 13-1*b*.

The substrate is usually a wafer of *N*-type silicon 150 to 200 μm thick with a resistivity of about 1 to 10 ohm·cm. The substrate contains two P^+-regions (the source and drain) with a surface hole concentration of 10^{18} to 10^{20} cm^{-3}, produced by local diffusion. The distance between the source and drain, that is, the length of the channel, is from 5 to 50 μm, and the depth of the P^+-regions is from 2 to 3 μm. These regions are connected by a thin (2 to 5 μm) surface

layer which constitutes the *P*-channel. The semiconductor crystal is given a coat of dielectric film (SiO_2) 0.8 to 2 μm thick. Above the channel, this film is coated by a metal film 5 to 7 μm thick serving as the gate. The same metal film applied in the windows of the SiO_2 layer above the source and drain makes their terminals.

Principle and Modes of Operation. Characteristics. Figure 13-8 shows a depletion-mode MOSFET connected in common-source circuits differing in gate voltage polarity. The circuit of Fig. 13-8*a* corresponds to the depletion-mode MOSFET operating in the *depletion*

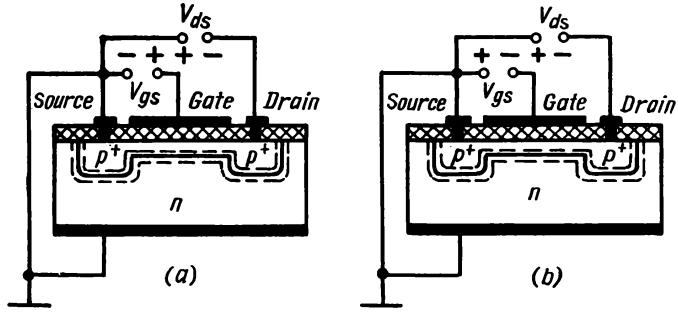


Fig. 13-8. Common-source arrangement of a depletion-mode MOSFET(*P*-channel)
(*a*) depletion mode; (*b* enhancement mode

mode, and that of Fig. 13-8*b* to the depletion-mode MOSFET operating in the *enhancement mode.* In the former case, the electric field developed by the gate potential which is positive with respect to the transistor substrate drives holes out of the channel, thereby depleting the majority carriers there. In the latter case, the electric field of the opposite polarity attracts holes into the channel, enhancing the majority carriers there. Thus, as with the JFET, the channel conductivity may be controlled by varying V_{gs} in magnitude and sign. The only difference is that in the depletion-mode MOSFET the concentration of charge carriers is varied in the device, rather than over the cross-sectional area of the channel.

Under the action of the potential difference between the source and drain ($-V_{ds}$), holes drift through the channel, thereby constituting the drain current, I_d. At the interface between the substrate and the P^+-regions and the channel, a *P-N* transition region is formed whose depletion layer is bounded by the dashed line in Fig. 13-8. The potential difference across the transition region varies from zero at the source to $-V_{ds}$ at the drain. Thus, the *P-N* transition region is reverse-biased along almost its full length, and there are no majority carriers moving across it. This holds for any polarity of the voltage at the gate which is insulated from the semiconductor crystal by a dielectric film. Owing to this, a depletion-mode MOSFET can be operated with either negative or positive gate voltage, whereas

in the JFET, the voltage between the gate and the semiconductor crystal (across the P-N junction) should always be reverse. Accordingly, the characteristics of the device are different (Fig. 13-9).

The Enhancement-Mode MOSFET. Physical Processes. An enhancement-mode MOSFET connected in a common-source circuit is

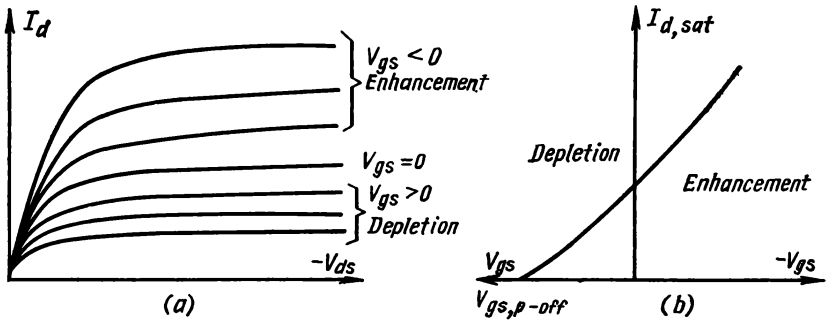


Fig. 13-9. (*a*) Output (drain) characteristics and (*b*) transfer (drain-gate) characteristic of a depletion-mode MOSFET (*P*-channel)

shown in Fig. 13-10. In contrast to the depletion-mode MOSFET, it operates only when a voltage negative relative to the source and substrate is applied to the gate. With zero gate voltage ($V_{gs} = 0$), there is no current flowing between the source and drain because the potential difference between the source and substrate is zero, and a reverse voltage, $-V_{ds}$, is applied between the drain and substrate.

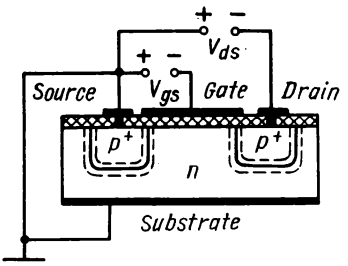


Fig. 13-10. Common-source configuration of an enhancement-mode MOSFET

At the interface between the substrate and the dielectric, there is a potential barrier: the subsurface layer of the N-type semiconductor crystal is charged negatively, and the dielectric film accumulates positive charges. This can be explained as follows.

Let us assume that the voltages at the transistor electrodes are equal to zero ($V_{gs} = 0$ and $V_{ds} = 0$). There are two P-N junctions, one each at the interface between the source and drain and the substrate, and a potential barrier at the interface between the dielectric and substrate. The potential barrier is formed because the surface states of N-type silicon at the boundary with the SiO_2 film are abundant in traps which capture electrons, whereas the SiO_2 film contains a certain amount of positive ions making up a surface charge Q_{surf} equal in magnitude to the negative charge at the surface of the semiconductor as shown in Fig. 13-11*a*. The left-hand side

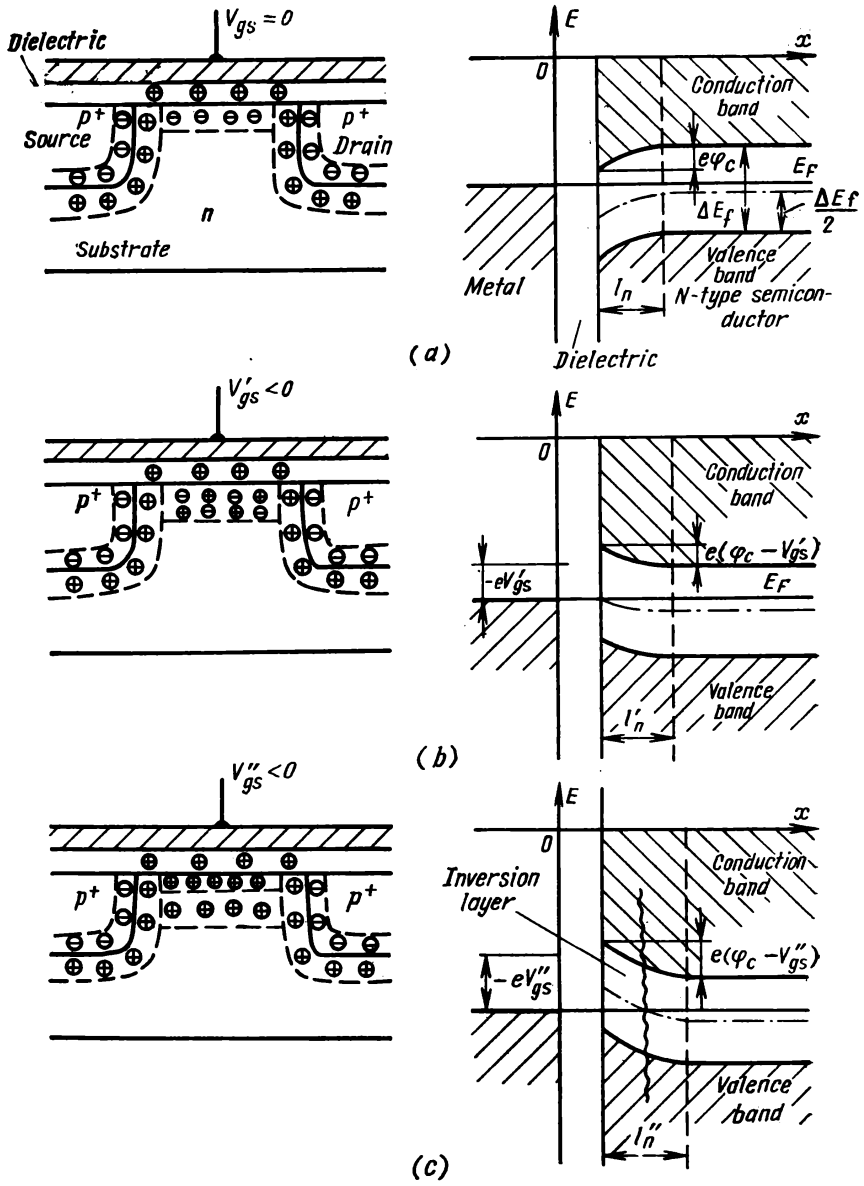


Fig. 13-11. Operating region and energy-band diagrams of an enhancement-mode MOSFET

(*a*) at $V_{gs} = 0$; (*b*) at low negative V_{gs}; (c) at negative V_{gs} at which the P-channel is formed

of this figure shows a part of the transistor structure near the drain and source regions. The right-hand side shows the energy-band diagram of the metal-dielectric-*N*-type-semiconductor junction.

The dipole layer forms a potential barrier, φ_c, at the surface of the semiconductor, and the boundaries of the energy bands of the semiconductor are shifted towards the region of negative potentials.

When a negative voltage is applied to the gate, the resulting electric field repels electrons from the surface into the bulk of the N-type semiconductor. As the gate voltage grows more negative, the potential barrier at first decreases to zero, and then, at $|V_{gs}| > |\varphi_c|$, changes its sign. At $-V_{gs} = -V'_{gs}$, the potential barrier is such that at the interface between the dielectric and semiconductor, the Fermi level is at the centre of the forbidden band (Fig. 13-11*b*). In this case, the surface layer of the N-type semiconductor has intrinsic conduction. As the gate voltage grows more negative, the surface layer of the semiconductor is depleted of electrons still more. The boundaries of the energy bands near the surface of the N-type semiconductor bend so much that in the narrow region near the surface the Fermi level is below the centre of the forbidden band. In other words, a thin layer with hole conduction is formed near the surface of the N-type semiconductor (Fig. 13-11*c*). Under the action of an electric field, this layer stores holes, and, as a result, a P-channel is formed between the drain and source. The channel which is a few thousandths of a micrometre thick is separated from the rest of the N-type semiconductor by a positive space charge established by the donor ions left behind by the electrons moving into the bulk of the semiconductor.

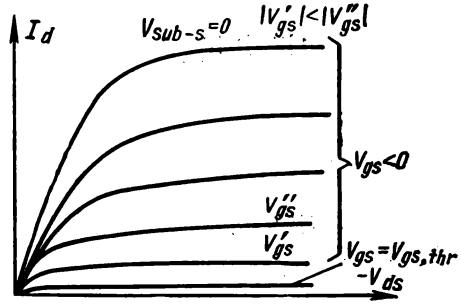


Fig. 13-12. Output (drain) characteristics of an enhancement-mode MOSFET

Varying the gate voltage, $-V_{gs}$, we may change the concentration of holes in the channel over a wide range, thereby controlling the channel conductivity. The value of the gate voltage at which the channel is formed in the transistor is known as the *threshold voltage*, designated $V_{gs,\,thr}$.

If, at $|V_{gs}| > |V_{gs,\,thr}|$, a negative voltage is applied to the drain ($V_{ds} < 0$), a longitudinal electric field appears in the channel, and holes begin to drift from the source to drain, thus making up the drain current I_d. This current may be controlled by varying $-V_{gs}$ (that is, the hole concentration in the channel) and V_{ds}. In the latter case, not only the drift velocity of holes in the channel and I_d will change, but also the shape of the channel. This we shall discuss in the study of the output characteristics.

Output Characteristics. These characteristics, $I_d = \varphi_2 (V_{ds})$ at V_{gs} = constant, are shown in Fig. 13-12.

They are practically the same as those for other types of FET. At first, I_d increases almost linearly with increasing negative V_{ds} and increasing longitudinal electric field in the channel. The characteristic departs from a straight line because the channel resistance varies as its cross-sectional area decreases with increasing distance from the source. When V_{ds} is applied in negative polarity, the *P-N* transition region between the drain and substrate is reverse-biased ($V_{sub\text{-}s} = 0$), while the source-to-substrate voltage remains equal

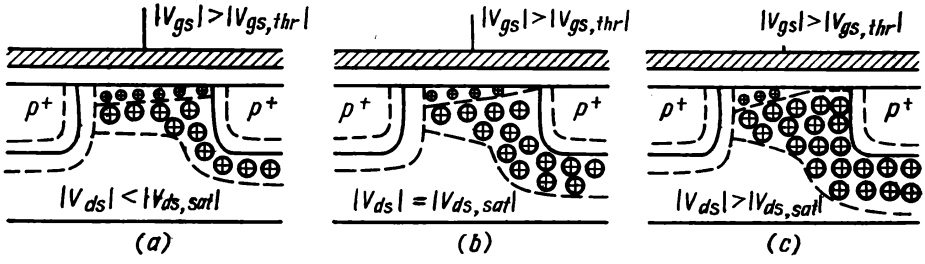


Fig. 13-13. Operating region at different V_{ds}
(a) at $|V_{ds}| < |V_{ds,sat}|$; (b) at $|V_{ds,sat}|$; (c) at $|V_{ds}| > |V_{ds,sat}|$

to zero. The depletion layer at the drain transition region widens, the positive space charge established by donor ions increases, and holes are partly swept out of the channel at its drain end. As a result, the channel narrows down (Fig. 13-13*a*).

At $|V_{ds}| = |V_{ds,sat}|$, the space charge grows still more positive, and the channel at the drain end becomes a narrow strip (Fig. 13-13*b*). The transistor is said to attain saturation. In the saturation region, I_d is insignificantly affected by variations in $|V_{ds}|$ (gradual portions of the characteristics). As $|V_{ds}|$ rises further, the beginning of the channel neck shifts towards the source (Fig. 13-13*c*), and the channel resistance increases still more. The channel resistance does not rise in proportion to $|V_{ds}|$, so I_d somewhat increases.

Analytically, the relation between I_d and the voltages is described by an approximate expression given in reference [24]:

$$I_d = g_{m,h} [(|V_{gs}| - |V_{gs,thr}|) |V_{ds}| - |V_{ds}^2|/2] \qquad (13\text{-}10)$$

where

$$g_{m,h} = \partial^2 I_d / \partial V_{gs} \partial V_{ds} \qquad (13\text{-}11)$$

Differentiating Eq. (13-10) with respect to V_{ds} and putting the derivative equal to zero, we get the saturation voltage:

$$|V_{ds,sat}| = |V_{gs}| - |V_{gs,thr}| \qquad (13\text{-}12)$$

Substituting this expression into Eq. (13-10), we may write for the drain current in the saturation region:

$$I_{d,sat} = (g_{m,h}/2)(|V_{gs}| - |V_{gs,thr}|)^2 \qquad (13\text{-}13)$$

The Effect of the Substrate. Figure 13-14 shows another family of drain characteristics. In contrast to Fig. 13-12, they are plotted for different substrate voltages, $V_{sub\text{-}s}$, while the gate voltage, $|V_{gs}|$, remains the same for the whole family of curves. Thus, these characteristics are given for a circuit configuration where the substrate

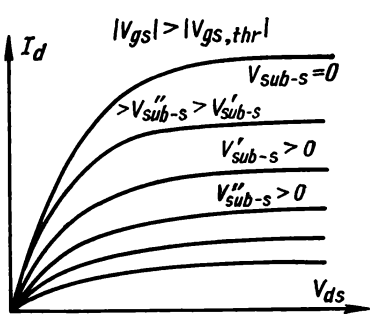


Fig. 13-14. Output (drain) characteristics of a FET at different substrate voltages

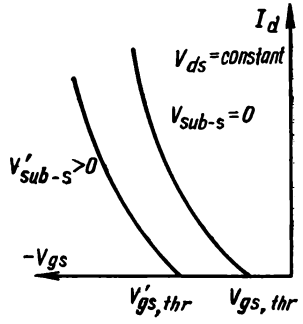


Fig. 13-15. Transfer (drain-gate) characteristics at different substrate voltages

is not connected to the source as in Fig. 13-10, but is held at a positive potential ($V_{sub\text{-}s} > 0$) with respect to the source. In this case, the substrate is used as an additional control electrode which governs the current through the channel. This is why the substrate is sometimes called the *lower gate*. The mechanism of current control is the same as in the JFET. The electric field developed by $V_{sub\text{-}s}$ not only widens the region near the channel, depleted of electrons, but also drives some holes out of the channel. The channel resistance increases and, as can be seen from the characteristics of Fig. 13-14, I_d decreases.

Transfer (Drain-Gate) Characteristic. This characteristic, $I_d = \varphi_1(V_{gs})$, is shown in Fig. 13-15. At $V_{sub\text{-}s} = 0$, the drain current begins to flow in the transistor only when $|V_{gs}| > |V_{gs,thr}|$, that is, after the channel has been formed. When a positive voltage is applied to the substrate, the drain gate characteristic shifts towards the more negative gate voltages. The channel is formed at $|V'_{gs,thr}| > |V_{gs,thr}|$.

13-4. FET PARAMETERS

The main parameters of field-effect transistors are the *y*-parameters in Eqs. (13-1) written for an equivalent two-port. In the general case, the currents and voltages of the FET are complex in nature, so use is made of admittances which for the common-source configuration are as follows

—input admittance

$$Y_{11s} = \frac{d\dot{I}_g}{d\dot{V}_{gs}}\bigg|_{V_{ds}=\text{constant}} \tag{13-14}$$

—reverse transfer admittance

$$Y_{12s} = \frac{d\dot{I}_g}{d\dot{V}_{ds}}\bigg|_{V_{gs}=\text{constant}} \tag{13-15}$$

—forward transfer admittance

$$Y_{21s} = \frac{d\dot{I}_d}{d\dot{V}_{gs}}\bigg|_{V_{ds}=\text{constant}} \tag{13-16}$$

—output admittance

$$Y_{22s} = \frac{d\dot{I}_d}{d\dot{V}_{ds}}\bigg|_{V_{gs}=\text{constant}} \tag{13-17}$$

The small-signal operation of field-effect transistors is described in terms of conductances, namely g_{11s}, g_{12s}, g_{21s} and g_{22s}.

In the literature of the subject these quantities are often designated differently than those for vacuum tubes.

The FET parameters of importance are listed below.

The *input dynamic resistance* is defined as

$$R_{in,dyn} = 1/g_{11s} = \frac{dV_{gs}}{dI_g}\bigg|_{V_{ds}=\text{constant}} \tag{13-18}$$

As already noted, $R_{in,\,dyn}$ for field-effect transistors is very high, being 10^8 to 10^{10} ohms for the JFET and 10^{10} to 10^{12} ohms for the IGFET.

The *transconductance* is defined as

$$g_m = g_{21s} = \frac{dI_{d,sat}}{dV_{gs}}\bigg|_{\substack{V_{ds}=\text{constant}\\ V_{sub-s}=0}} \tag{13-19}$$

The transconductance ranges between several tenths of a milliampere per volt and 3 mA/V.

The relation between the transconductance and the voltages may be derived by differentiating Eqs. (13-8) and (13-10). For the JFET

$$g_m = (1/r_{ch,dyn0})\,(1 - \sqrt{|V_{gs}|/|V_{gs,p-off}|} \tag{13-20}$$

For the IGFET

$$g_m = g_{m,h}(|V_{gs}| - |V_{gs,thr}|) \tag{13-21}$$

The *output dynamic resistance* (the drain dynamic resistance) is defined as

$$R_{id} = R_{out,dyn} = 1/g_{22s} = \frac{dV_{ds}}{dI_{d,sat}}\Big|_{V_{gs}=\text{constant}} \tag{13-22}$$

The value of R_{id} is rather high, being from 10^5 to 10^7 ohms.

The effect of the substrate voltage on I_d is expressed in terms of the substrate transconductance which is defined as

$$g_{m,sub} = \frac{dI_{d,sat}}{dV_{sub}}\Big|_{\substack{V_{gs}=\text{constant}\\ V_{ds}=\text{constant}}} \tag{13-23}$$

As a rule, $g_{m,\,sub}$ is smaller than g_m. When the gate is connected to the substrate, the total transconductance is $g'_m = g_m + g_{m,\,sub}$.

By analogy with the static (d.c.) parameters for vacuum tubes, use is sometimes made of the parameter similar to the static amplification factor

$$M = g_m R_{id} = -\frac{dV_{ds}}{dV_{gs}}\Big|_{I_{d,sat}=\text{constant}} \tag{13-24}$$

The performance of the FET as an amplifier, radio signal converter, and also in pulse work is considerably affected by interelectrode capacitances. In field-effect transistors, there is neither injection of minority carriers through the *P-N* junction nor their diffusion in the base, which affects the frequency properties of bipolar transistors. The main processes responsible for the time lag in the FET and, as a consequence, its frequency properties, are the charging and discharging of the interelectrode capacitances.

The capacitances of importance to the performance of the FET are the gate-source capacitance, C_{gso}, measured with the other leads open-circuited for alternating current; the input gate-source capacitance, C_{11s}, measured with the output short-circuited for alternating current; the output drain-source capacitance, C_{22s}, measured with the input short-circuited for alternating current; and the transfer gate-drain capacitance, C_{12s}, measured with the input short-circuited for alternating current.

In field-effect transistors, the interelectrode capacitances are equal to several tenths of a picofarad.

Pulse voltage at the output of the FET is described in terms of the time intervals defining the rise and fall of the output pulse, namely the ON delay time, $t_{d,\,ON}$; the rise time, t_r; the OFF delay time, $t_{d,\,OFF}$; and the fall time, t_f. These parameters are similar

to those for bipolar transistors (see Sec. 12-7), except that t_s is replaced by $t_{d,\,OFF}$. This is because in field-effect transistors, there is no minority carrier dissipation in the base characteristic of bipolar transistors. The above time intervals are measured similarly to those of the bipolar transistor, that is, between 0.1 and 0.9 of the peak (final) value of the output pulse.

Chapter Fourteen

PHOTOELECTRIC AND LIGHT-EMITTING SEMICONDUCTOR DEVICES

14-1. GENERAL

Definition. *Semiconductor photoelectric* devices are those intended to convert the energy of incident light into electric energy.

As their name implies, *semiconductor light-emitting devices* are those converting electric energy directly into light.

For their operation, photoelectric devices depend on the photoconductive and photovoltaic effects which will be discussed later in this chapter.

Classification. Semiconductor photoelectric devices are classified into *photoresistive* and *photovoltaic cells.* The former include *photoresistors*, *photodiodes*, *phototransistors* and *photothyristors.* The latter consist of various photocells including *solar batteries.*

The physical quantities characterizing radiant energy are briefly discussed in Sec. 6-1.

In the Soviet Union, the diagram symbols for semiconductor photoresistive and photovoltaic cells are established by USSR State Standards GOST 2.730-68, GOST 2.729-68 and GOST 17704-72. The nomenclature and definitions are set up by GOST 15133-69 and GOST 19852-74.

14-2. ABSORPTION OF LIGHT BY SEMICONDUCTORS

Laws of absorption. When a semiconductor crystal is exposed to light, it absorbs some luminous energy and reflects the rest of it. The reflection of a luminous flux by a semiconductor surface is expressed in terms of the *reflectance*

$$R = \Phi_r/\Phi \tag{14-1}$$

where Φ and Φ_r are the incident and reflected luminous fluxes, respectively.

The absorption of light by a semiconductor is accompanied by a decrease in the luminous flux in accordance with the law known from physics

$$\Phi(x) = \Phi(1 - R) \exp(-\alpha_0 x) \tag{14-2}$$

where α_0 is the *absorptance* showing the absorbed energy per unit length and measured in cm^{-1}, and x is the thickness of the semiconductor crystal.

According to their structure, the absorption of light by semiconductors may be accompanied by changes in the energy states of free or valence electrons, impurity atom electrons, or in the crystal lattice itself. Thus, absorption centres in solids vary in nature. According to the prevalent absorption centres, substances absorb electromagnetic waves of different wavelengths unequally. The relation between the absorptance and the wavelength of incident light, $\alpha_0(\lambda)$, is called the *absorption spectrum.*

Naturally, the absorption of light quanta (photons) by a crystal should obey the law of conservation of energy and momentum. To demonstrate, if a photon has the energy $h\nu$ and momentum P_{ph}, and an electron before absorption has the energy E_{e1} and the momentum p_{e1}, then, after absorption, they should be as follows

$$E_{e2} = E_{e1} + h\nu \tag{14-3}$$

$$p_{e2} = p_{e1} + p_{ph} \tag{14-4}$$

This is accompanied by various physical processes which markedly depend on the energy-band structure of a semiconductor. Now, we shall discuss the main types of light absorption by semiconductors.

Intrinsic Absorption. One of the main types of light absorption — *intrinsic* or *fundamental absorption*—is associated with the transition of an electron from the valence into conduction band. This transition is possible if the energy of the absorbed photon exceeds or is, at least, equal to the forbidden band width, that is, $h\nu \geqslant \Delta E_f$. As has been shown in Chapter 9, semiconductors differ in energy-band structure. In some semiconductors, the bottom of the conduction band and the top of the valence band correspond to the same quasimomentum or wave vector $k = 2\pi p/h$. An illustrative example of such semiconductors is furnished by indium antimonide, InSb (see Fig. 9-9*b*). In most semiconductors, however, including germanium and silicon, these energy extrema correspond to different k (see Fig. 9-9*c*).

In the first group, the transition of an electron from band to band caused by intrinsic absorption results only in a change in its energy. The electron momentum remains practically unchanged. Such transitions are called *direct* or *vertical.* Owing to the absorption of a photon in a direct transition, the change in the electron momentum caused by the photon momentum, $p_{ph} = h/\lambda_{ph}$, may be neglected because the photon wavelength is λ_{ph} = (approx.) 10^{-1} to 10^{-5} cm and the wavelength of the electron wave function at T = (approx.) 300 K is λ_e = (approx.) 5×10^{-7} cm. Consequently, the electron momentum is much greater than the photon momentum, and

Eq. (14-4) may be re-written as follows

$$p_{e2} = (\text{approx.})\, p_{e1} \tag{14-5}$$

In the second group of semiconductors, the transition of an electron from the valence into conduction band should be accompanied not only by a change in its energy but also by a change in its momentum. This is what is known as *indirect transition.* In an indirect transition, the absorption of a photon is complemented by the absorption of a phonon whose energy is usually small (being several tenths of an electron-volt), and the momentum, p_{phn}, may reach a considerable value, so Eq. (14-4) takes the form

$$p_{e2} = p_{e1} + p_{phn} \tag{14-6}$$

As is seen, indirect transitions are associated with changes in the energy of the crystal lattice. Obviously, the energy exchange with the crystal lattice is subject to additional requirements which cannot be satisfied always. Because of this, the absorptance α_0 for indirect transitions ($\alpha_0 \approx 10^{-1}$ to 10^3 cm^{-1}) is usually below that for direct transitions ($\alpha_0 \approx 10^4$ to 10^5 cm^{-1}).

Intrinsic absorption in a semiconductor generates pairs of mobile electric charges, namely holes in the valence band and electrons in the conduction band. The probability that the carriers will move from band to band, that is, the probability that an electron-hole pair will be formed, depends on the occupancy of the energy states near the top of the valence band, and also on the density of vacant energy states near the bottom of the conduction band.

The intrinsic absorption spectrum is sufficiently wide, because the absorption of photons having an energy exceeding the width of the forbidden band enables electrons to move into higher energy levels farther away from the bottom of the conduction band. These free electrons having a higher energy (sometimes called *hot electrons*) dissipate the excess energy on the crystal lattice during a certain mean time t_d and sink to lower vacant energy levels at the bottom of the conduction band. As a rule, t_d is much shorter than the average lifetime of a free charge carrier measured between its generation and recombination.

On the longwave side, the boundary of the intrinsic absorption spectrum is determined by ΔE_f, the minimum photon energy which may cause an interband transition. In most semiconductors, intrinsic absorption takes place in the visible or near infrared regions of the spectrum, depending on the width of the forbidden band.

Exciton Absorption. When a semiconductor absorbs a photon, its valence electron may be excited in such a way that, instead of moving to a higher energy state, it forms a bound electron-hole pair known as an *exciton.* In an exciton, the electron and hole are held

together by Coulomb forces whose magnitude depends on the dielectric constant, ε_0, and relative permittivity of the crystal, ε. The exciton is customarily regarded as a particle with an effective mass $M^* = m_n^* + m_p^*$.

Associated with an exciton is a set of energy eigenvalues which form a hydrogen-like energy spectrum lying within the forbidden band of the semiconductor. So, the exciton absorption spectrum of a semiconductor contains a number of extrema corresponding to the generation of an exciton with an energy of one of the possible states.

Extrinsic Absorption. This type of light absorption is associated with the ionization or excitation of impurity atoms. The absorption of photons causes donor electrons to move into the conduction band, or forces the valence electrons of the semiconductor to move to acceptor levels. In addition, the electrons of impurity centres may move to the energy levels associated with the excitation of these atoms. For the most commonly used dopants, the ionization energy of the donor atoms is from a few tenths to a few hundredths of the forbidden band width and is usually equal to a few hundredths of an electronvolt. So, the extrinsic absorption spectrum usually lies beyond the longwave boundary of intrinsic absorption. The extrinsic absorption spectra cover a wide range of frequencies, because during light absorption, donor electrons may move to vacant energy levels in the conduction band sufficiently far away from its bottom, and acceptor atoms may be ionized by electrons coming from the deeper energy levels of the valence band.

As with the thermal ionization of impurity atoms, extrinsic absorption results in the generation of mobile charge carriers of only one polarity, namely electrons in the conduction band which are released due to the ionization of donor atoms, and holes in the valence band produced by the ionization of acceptor atoms. The impurity atoms themselves are turned in the course of photoionization into ions with unit electric charge.

Obviously, the absorptance α_0 in extrinsic absorption markedly depends on temperature. At room temperature, almost all shallow impurity levels are thermally ionized, so the probability of extrinsic absorption is low; at a lower temperature, the intensity of extrinsic absorption increases.

Other Types of Absorption. In addition to intrinsic and extrinsic absorption, photons may be absorbed by free charge carriers and the crystal lattice itself. In the former case, free charge carriers—electrons in the conduction band and holes in the valence band—absorb photons and move to higher vacant energy levels. Such transitions are indirect and, as a consequence, are accompanied by the absorption of a photon. Here, the absorptance is proportional to the number density of charge carriers and the squared wavelength of incident light, and inversely proportional to their mobility, μ,

and effective mass. In this case, α_0 is anywhere between 100 and 500 cm^{-1}.

In semiconductors having a complex energy-band structure, for example, silicon (Fig. 9-1*c*), the absorption of photons enables charge carriers to move from valley to valley, and this, as will be recalled, may change their effective mass.

The absorption of light by the crystal lattice is related to the interaction of electromagnetic energy in the form of light waves with the oscillating lattice vertices. In this case, the energy of photons is converted to the energy of phonons. Since such an interaction is resonant in nature, the absorption spectrum here consists of a number of absorption peaks corresponding to different wavelengths of incident light.

As is seen, the absorption of light by semiconductors is associated with various physical processes in the crystal. The absorption may cause the generation of additional free charge carriers and, as a consequence, a change in the resistance of the crystal (this is known as the *photoconductive effect*).

Illumination of a semiconductor crystal containing a *P-N* transition region may give rise to an electromotive force between the two regions or between the semiconductor and metal (this is the *photovoltaic effect*). These and other physical effects underlie the operation of various photoelectric semiconductor devices.

14-3. THE PHOTOCONDUCTIVE EFFECT

Definition. The photoconductive effect refers to the change in the resistivity (or conductivity) of a semiconductor caused solely by electromagnetic radiation and not connected with its heating.

Photoconductivity. The general features of the photoconductive effect are that the conductivity of a material increases with illumination by a certain value known as the *photoconductivity*.

Let us designate the number densities of equilibrium carriers in an intrinsic semiconductor as n_0 and p_0. Then, its intrinsic conductivity (in the absence of illumination) is

$$\sigma_0 = e(n_0\mu_n + p_0\mu_p) \tag{14-7}$$

This quantity is sometimes called the *dark conductivity*.

When a semiconductor is illuminated, the concentrations of mobile charge carriers change due to intrinsic absorption to $n_0 + \Delta n$ and $p_0 + \Delta p$, and its conductivity is

$$\sigma = e[(n_0 + \Delta n)\mu_n + (p_0 + \Delta p)\mu_p] \tag{14-8}$$

The difference between the conductivity σ and the dark conductivity σ_0 gives the photoconductivity of the semiconductor:

$$\sigma_{ph} = \sigma - \sigma_0 = e(\Delta n\mu_n + \Delta p\mu_p) \tag{14-9}$$

In the case of extrinsic absorption, either positive or negative nonequilibrium carriers prevail in the semiconductor. To demonstrate, if absorption is due to the thermal ionization of donor impurities, then $\Delta n \gg \Delta p$, and, as a consequence, the extrinsic photoconductivity is

$$\sigma_{ph,ex} = (\text{approx.})\, e\Delta n\mu_n \qquad (14\text{-}10)$$

The number of electron-hole pairs generated on the unit illuminated surface area is proportional to the number of photons per second, N, the absorptance, α_0, and the quantum efficiency, β, giving the number of electron-hole pairs generated by one absorbed photon

$$\Delta n = \Delta p = \alpha\beta N \qquad (14\text{-}11)$$

When a semiconductor is struck by light, the rate of electron-hole pair generation does not rise to its steady-state value corresponding to the intensity of incident light at once, but does so exponentially

$$\Delta n(t) = \alpha\beta\tau N\,[1 - \exp(-t/\tau)] \qquad (14\text{-}12)$$

where τ is the lifetime of nonequilibrium carriers. At $t \gg \tau$, the nonequilibrium carriers number density reaches a steady-state value

$$\Delta n_{st} = \alpha\beta\tau N \qquad (14\text{-}13)$$

When light ceases to illuminate the semiconductor, Δn decreases exponentially*

$$\Delta n(t) = \Delta n_{st} \exp(-t/\tau) = \alpha\beta\tau N \exp(-t/\tau) \qquad (14\text{-}14)$$

The photoconductivity varies in accordance with the rise and fall in the concentration of nonequilibrium carriers. These gradual changes in σ_{ph} when the illumination is turned on and off are called the *relaxation of photoconductivity*.

The relaxation time may be markedly affected by surface recombination, trap recombination, and also the diffusion of nonequilibrium carriers into the bulk of the semiconductor crystal.

The *spectral photoconductivity characteristic*, that is, a plot of photoconductivity σ_{ph} as a function of the wavelength of incident light, is shown in Fig. 14-1. Portion *1* of the curve corresponds to intrinsic absorption whose longwave boundary in micrometres may be found on the basis of the following relation

$$\lambda''_{in} = hc/\Delta E_f = (\text{approx.})\, 1.24/\Delta E_f \qquad (14\text{-}15)$$

where c is the velocity of light.

* Equations (14-12) and (14-14) hold only when $\Delta n \ll n_0 + p_0$. Otherwise, Δn rises in accordance with the hyperbolic tangent law, and falls in accordance with the hyperbola law given in reference [22].

The boundary wavelength of extrinsic photoconductivity, λ''_{ex}, (portion *2*) in micrometers is

$$\lambda''_{ex} = (\text{approx.})\, 1.24/\Delta E_d = (\text{approx.})\, 1.24/\Delta E_a \qquad (14\text{-}16)$$

Here, ΔE_d and ΔE_a are the ionization energies of donors and acceptors, respectively, which, as well as ΔE_f in Eq. (14-15), are expressed in electronvolts.

14-4. PHOTORESISTORS

Definition. *Photoresistors* are bipolar semiconductor resistors which, when illuminated, drop in resistance.

Photoresistors are used to form electric signals when struck by light signals, which may be constant in time (unmodulated signals)

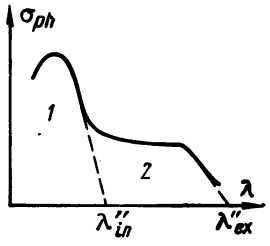


Fig. 14-1. Spectral response of a photoconductor

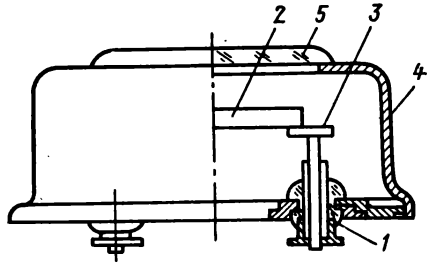


Fig. 14-2. Construction of a photoresistor

1—insulator; *2*—semiconductor; *3*—metal terminals; *4*—case; *5*—mica or glass window

or vary sinusoidally or according to any other law (modulated signals).

Also, photoresistors are used to detect light signals. In this case, photoresistors, as well as photodiodes and phototransistors (see Secs. 14-7 and 14-8), serve as sensing units at receiver inputs in optical communication systems, infrared detection systems, radio astronomy equipment, and so on.

Design. Figure 14-2 shows the construction of a photoresistor. The light-sensitive semiconductor wafer is made either by pressing a semiconductor powder or by depositing this material on a dielectric base in a vacuum. In some cases, the wafer is made of a single semiconductor crystal. On the surface of the semiconductor wafer are deposited thin-film metal contacts connected to leads. The surface of the wafer between the metal contacts constitutes the effective area of the photoresistor, which ranges between 0.5 and 30 mm^2 from device to device.

Most commonly, use is made of type $A^{II}B^{VI}$ (CdSe, CdS) or $A^{IV}B^{VI}$ (PbS, PbSe) semiconductors.

Characteristics and Parameters. The *volt-ampere characteristic* (Fig. 14-3) shows the current through a photoresistor as a function of the applied voltage at different values of the luminous flux. In darkness, the photoresistor is conducting because of free charge carriers, both electrons and holes, formed due to thermal excitation. At a certain operating voltage, V_{op} and $\Phi = 0$, the current through the photoresistor is called the *dark current*, designated I_d, and at $\Phi > 0$, it is termed the *total current*, labelled I_{tot}. The difference between the two currents is equal to the *photocurrent*:

$$I_{ph} = I_{tot} - I_d \qquad (14\text{-}17)$$

The illumination dependence of photocurrent (Fig. 14-3*b*) is linear at low Φ. As Φ increases, the photocurrent grows at a slower rate,

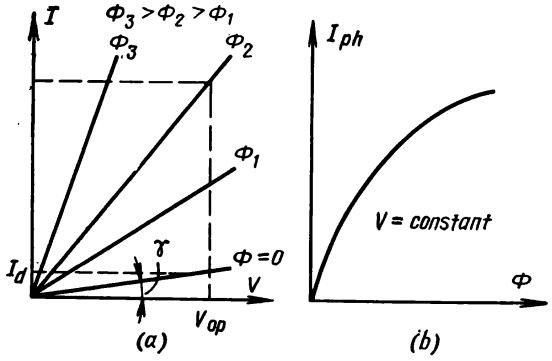


Fig. 14-3. (*a*) Volt-ampere and (*b*) current-illumination characteristics of a photoresistor

because a rise in the concentration of free charge carriers raises the probability of their trap recombination, see reference [22], and their lifetime τ shortens.

Sometimes, this characteristic is termed the *lux-ampere characteristic*, the abscissa being illumination in luxes rather than luminous flux.

Figure 14-4 shows the relative spectral characteristic. The ordinate is monochromatic sensitivity, s_λ, to be defined later. The figure shows the spectral characteristics of materials whose sensitivity reaches its peak at intrinsic absorption. In addition, photoresistors are also made of materials with maximum sensitivity at extrinsic absorption. Their spectral characteristics will be discussed later in this section.

Referring to Fig. 14-4, due to different ΔE_f, maximum sensitivity for various materials occurs in different portions of the light spectrum, from the yellow-green region for cadmium sulphide photoresistors to the near infrared region for lead sulphide and lead selenide devices.

One of the major parameters for any photoelectric device is its *sensitivity*. For photoresistors, use is most commonly made of *current sensitivity*, s_I, which is defined as the ratio of the photocurrent to a certain quantity characterizing the radiation which caused the photocurrent being measured*. If, for example, this quantity is the luminous flux, we speak of the *luminous sensitivity* of a photoresistor, s_Φ. There are also *radiant sensitivity*, s_F, *illumination sensitivity*, s_E, etc. If the sensitivity is determined by the ratio of two constant values, it is called *static sensitivity*, s_{st}; in the case of incremental changes in these quantities, it is termed *dynamic sensitivity*, s_{dyn}.

Sensitivity varies with the spectrum of incident light. The sensitivity of a device to nonmonochromatic radiation is known as *integrated sensitivity*, s_{int}, and that to monochromatic radiation is called *monochromatic sensitivity*, s_λ.

The performance of photoresistors is expressed in terms of integrated specific luminous sensitivity measured in mA/(V lm):

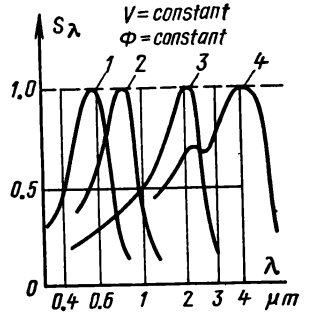


Fig. 14-4. Relative spectral response of various photoresistors

1—cadmium sulphide; *2*—cadmium selenide; *3*—lead sulphide; *4*—lead selenide

$$s_{\Phi,int,sp} = I_{ph}/\Phi V \qquad (14\text{-}18)$$

This sensitivity is called specific because it gives the integrated sensitivity per volt of the applied voltage. For industrial photoresistors, $s_{\Phi,int,sp}$ varies from a few tenths to several hundred mA/(V lm). The integrated specific sensitivity is measured at an illumination of 200 lx.

Another parameter is the *dark resistance*, R_d, that is, the resistance measured at $\Phi = 0$. It can easily be determined from cot γ, where γ is the angle made by the dark volt-ampere characteristic of the photoresistor and the x-axis (Fig. 14-3*a*). The dark resistance ranges between several units and a few tens of megohms.

A sinusoidal signal modulating the incident luminous flux is described in terms of the *limiting frequency*, f_{lim}, defined as the signal frequency at which the sensitivity of a photoresistor falls to $1/\sqrt{2}$ of its sensitivity with an unmodulated flux. For most photoresistors, f_{lim} = (approx.) 10^3 to 10^4 Hz.

Another parameter is the *threshold flux* which is very important for photoelectric devices, including photoresistor, used to detect small light signals. The threshold flux, Φ_{thr}, is defined as the rms sinusoidally modulated flux with a specified spectral distribution

* The performance of photoelectric devices whose output is the voltage of a photosignal is described in terms of *voltage sensitivity*, s_v.

at which the rms output voltage or current of the detector is equal to the rms noise voltage or current in the specified frequency band.

In other words, the threshold flux is the minimum incident flux that can be detected against the background of detector noise.

The *threshold flux per hertz* is called the *unit bandwidth threshold flux* and designated Φ_{thr1}. The quantity $\Phi^*_{thr} = \Phi_{thr1}/s$, where s is the surface area of the photodetector, is called the *specific threshold flux*. More often, however, the performance of photodetectors is expressed in terms of *detectability* which is the reciprocal of the threshold luminous flux, $D = 1/\Phi_{thr}$, or in terms of their specific detectability, $D^* = 1/\Phi^*_{thr}$. This parameter is measured in W^{-1} cm $Hz^{1/2}$.

As with most semiconductor devices, the characteristics and parameters of photoresistors are markedly affected by temperature variations. This is customarily expressed in terms of the *temperature coefficient of photocurrent* defined as

$$\alpha_T = \frac{1}{I_{ph}} \frac{\partial I_{ph}}{\partial T} \bigg|_{\Phi = \text{constant}} \qquad (14\text{-}19)$$

This coefficient ranges between approximately -10^{-3} and -10^{-4}°C.

The normal *operating conditions* for a photoresistor are those which do not cause irreversible changes in its parameters during the entire service life. They are specified as ratings which include the operating voltage (between 2 and 100 V according to resistor type) and the maximum power dissipation (which runs into several milliwatts).

Types of Photoresistors. The photoresistors used to form electric signals when exposed to a radiant flux are made of cadmium sulphide, cadmium selenide, lead sulphide and lead selenide. The spectral characteristics of these devices are shown in Fig. 14-4. Cadmium-sulphide photoresistors have a high operating voltage (up to 400 V), significant maximum current (a few milliamperes), and very high integrated sensitivity reaching several amperes per lumen. The dark resistance of such devices is also very high, being from 10^7 to 10^{10} ohms.

Cadmium-selenide photoresistors also have a high integrated sensitivity (15 to 20 A/lm) and high dark resistance ($R_d \approx 5 \times \times 10^6$ ohms), but can handle lower operating voltages and currents.

Lead-sulphide, and lead-selenide photoresistors are very sensitive to infrared radiation. When the temperature of a device falls, its spectral characteristic shifts towards the longer waves. So, these photoresistors are used both at room temperature and also at the temperature of liquid nitrogen (77 K) when used in suitable coolers. The integrated sensitivity of such devices in the infrared region reaches 2 mA/W.

In the infrared region, use is made of photoresistors fabricated from germanium doped with gold, zinc, antimony, etc., and cooled

to the temperature of liquid nitrogen. Their spectral characteristic corresponds to extrinsic photoconductivity. Their specific detectability ranges between 10^{10} and 4×10^{10} W^{-1} cm $Hz^{1/2}$.

14-5. THE PHOTOVOLTAIC EFFECT

Definition. The *photovoltaic effect* refers to the electromotive force generated at metal semiconductor and semiconductor junctions upon excitation by electromagnetic radiation.

Accordingly, photovoltaic cells are self-generating semiconductor devices which convert light energy into electrical energy when struck by light. They include semiconductor photocells and also photodiodes, phototransistors and photothyristors (see Secs. 14-6, 14-7 and 14-8).

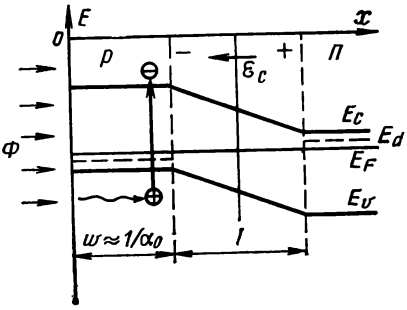


Fig. 14-5. Energy-band diagram of an illuminated P-N junction

Illumination of a P-N Junction. Figure 14-5 shows the energy-band diagram of a *P-N* junction exposed to a luminous flux Φ.

Let us assume that practically all of the impurity atoms are thermally ionized, and only intrinsic absorption takes place when the *P-N* junction is struck by light. Also we assume that Φ contains spectral components for which

$$h\nu \geqslant \Delta E_f \tag{14-20}$$

As will be recalled, the absorptance, α_0, depends on frequency. Suppose that at the frequencies satisfying the inequality (14-20), α_0 is sufficiently high so that absorption mainly occurs near the surface of the *P*-type semiconductor. Intrinsic absorption results in the generation of free electron-hole pairs. Thus both positive and negative carriers in the *P*-type semiconductor are present in a nonequilibrium concentration and diffuse towards the barrier region where a lower concentration exists. On reaching this region, electrons are captured by the junction field $\mathcal{E}_j$ and transferred to the *N*-region where they act as majority carriers. Holes are retarded by the junction field and remain in the *P*-region. As a result, the concentration of majority carriers increases on either side of the barrier region.

If the semiconductor regions making up the junction are not connected by an external circuit, the *P*-region accumulates holes and the *N*-region electrons. The space charge built up by these majority carriers balances some of the charge due to the ionized impurity atoms in the barrier region, and the potential barrier in

the junction decreases. The condition of equilibrium is upset, and a diffusion current due to majority carriers flows through the junction. A new state of equilibrium will be attained at a lower potential barrier equal to $\varphi_c - V_{ph}$, when the current of minority carriers generated by the absorption of photons will be fully balanced by the opposite diffusion current of majority carriers.

The resulting potential difference, V_{ph}, by which the potential barrier of the junction drops is called the *photoelectromotive force* (abbreviated *photo emf*).

The photoelectromotive force depends on the intensity of incident light and some other factors, but its maximum value cannot be higher than the contact potential difference, φ_c.

If we connect the semiconductor regions forming the *P-N* junction by an external circuit, it will carry the photocurrent, I_{ph}, constituted by the flow (through the junction) of minority carriers generated by photons, this current being in the same direction as the reverse current.

The number of electron-hole pairs generated per unit time is proportional to the number of photons absorbed, $\Phi/h\nu$, and the quantum efficiency, β. So, we may write for the photocurrent

$$I_{ph} = (e\Phi/h\nu)\,\beta\varkappa \qquad (14\text{-}21)$$

where $\varkappa$ is the carrier collection coefficient giving the number of electron-hole pairs that do not recombine in the bulk and at the surface of the semiconductor and reach the barrier region ($\varkappa \leqslant 1$).

Referring to the above equation, the photocurrent is proportional to the luminous flux. The constant of proportionality in Eq. (14-21) defined as

$$s_{int} = e\beta\varkappa/h\nu \qquad (14\text{-}22)$$

is the *integrated sensitivity* of the photovoltaic cell.

The coefficient $\varkappa$ depends on a number of factors, including the electrophysical properties of the material, the thickness of the semiconductor layer near the surface, and so on. The point is that the photocurrent depends on the motion of nonequilibrium "light" charge carriers. This motion includes the diffusion and drift of carriers towards and through the barrier layer, respectively, under the action of $\mathscr{E}_j$. Obviously, if absorption mainly occurs in the *P*-region, then for $\varkappa$ to be close to unity, it is necessary that the thickness of that region should be a fraction of the diffusion length ($w \ll L$). On the other hand, if w were reduced too much, absorption would mainly occur close to the surface where the probability of recombination at the surface levels is considerably higher.

The diffusion and drift of carriers towards and through the *P-N* junction introduce a certain time lag in photovoltaic cells.

In some devices, absorption takes place not only in the semiconductor region adjacent to the surface but also in the barrier region.

According to the analysis carried out in reference [23], in either case, the optimal relation between the quantum efficiency and response time of a photovoltaic cell is achieved when

$$w = \text{(approx.)}\ 1/\alpha_0 \tag{14-23}$$

The Volt-Ampere Characteristic of an Illuminated P-N Junction. Assume that the external circuit of a P-N junction is completed via a load resistor R_L connected to the metal coating on the N-type semiconductor side and to the semitransparent metal film on the surface of the P-type semiconductor (Fig. 14-6). The photocurrent, I_{ph}, due to minority carriers flowing through the junction is in the same direction with the reverse current, I_r, in the P-N junction. Across R_L, it develops a voltage drop, V_R (applied in positive polarity to the P-type semiconductor), which reduces the potential barrier and, as a consequence, gives rise to a forward current opposing I_{ph}.

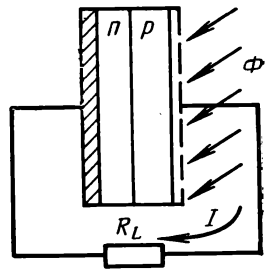


Fig. 14-6. Connection of a load resistor to a P-N junction

Recalling Eq. (10-52) for the volt-ampere characteristic of a P-N junction, we may write for the current in the external circuit of an illuminated P-N junction

$$I = I_{ph} - I_r\,[\exp(eV/kT) - 1] \tag{14-24}$$

Taking into consideration that $V = V_R = IR_L$, we may write

$$I = I_{ph} - I_r\,[\exp(eIR_L/kT) - 1] \tag{14-25}$$

Solving the above expression for V_R gives

$$V_R = (kT/e)\log_e[1 + (I_{ph} - I)/I_r] \tag{14-26}$$

These expressions describe the volt-ampere characteristic of an illuminated P-N junction (Fig. 14-7).

The forward (diffusion) slope of the volt-ampere characteristic of an unilluminated ($\Phi = 0$) forward-biased junction lies in the first quadrant. When light strikes the junction, the resulting photocurrent is subtracted from the diffusion current, and this part of the characteristic changes in shape. The family of curves depicting the behaviour of the reverse-biased P-N junction lies in the third quadrant. At $\Phi > 0$, the photocurrent is combined with the reverse current through the junction. This part of the characteristic can be recognized as a family of *displaced rectifier characteristics*. The curves lying in the fourth quadrant are known as the *semiconductor photocell characteristics*.

When the external circuit is shorted ($R_L = 0$, Fig. 14-6), the current in the external circuit is equal to the photocurrent and Eq. (14-26) takes the form

$$I_{s-c} = I_{ph} = s_{int}\Phi \tag{14-27}$$

When $\Phi_3 - \Phi_2 = \Phi_2 - \Phi_1 = \Delta\Phi$ (see Fig. 14-7), the curves make equal intercepts on the current axis.

In the open-circuit condition ($I = 0$), Eq. (14-26) takes the form

$$V_{o-c} = (kT/e)\log_e(1 + s_{int}\Phi/I_r) \tag{14-28}$$

The voltage across a photovoltaic cell rises logarithmically with increasing illumination. So, the intercepts made by the curves on

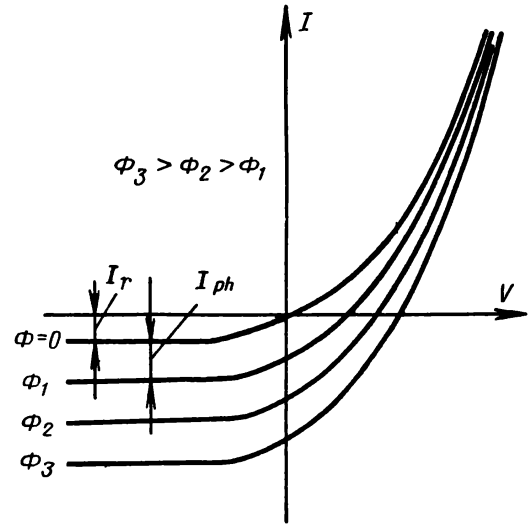


Fig. 14-7. Volt-ampere characteristics of an illuminated P-N junction

the voltage axis change in a similar manner.

As already noted, the voltage across an open-circuited photovoltaic cell gives the photoelectromotive force, V_{ph}. Substituting the values of constants into Eq. (14-28) gives for $T = 300$ K

$$V_{ph} = (\text{approx.})\,0.025\log_e(1 + s_{int}\Phi/I_r) \tag{14-29}$$

When illumination is strong, then $s_{int}\Phi \gg I_r$, and the unity in the brackets may be neglected. So,

$$V_{ph} = (\text{approx.})\,0.025\log_e(I_{ph}/I_r) \tag{14-30}$$

14-6. PHOTOELECTRIC CELLS

Definition. *Photoelectric cells* (or, simply, *photocells*) are light-sensitive devices utilizing the photovoltaic effect. They translate variations in light into corresponding variations in electrical signals, and are used to supply electronic circuits, in relay systems, etc.

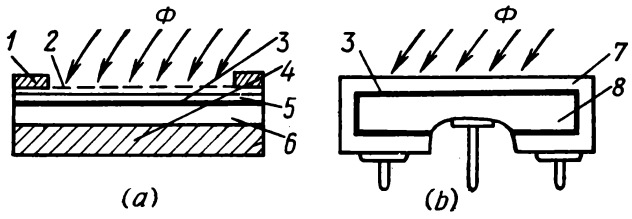


Fig. 14-8. Construction of semiconductor photocells

(*a*)—selenium; (*b*)—silicon; *1*—ring terminal; *2*—semitransparent metal layer; *3*—*P-N* junction; *4*—metal; *5*—*N*-type selenium; *6*—*P*-type selenium; *7*—*P*-type silicon; *8*—*N*-type silicon

Design. Figure 14-8 shows a sketch of a selenium and a silicon photocell. In the former, a layer of *P*-type selenium is vacuum-deposited on a substantial metal plate 1 to 2 mm thick which is then heated to $T = 200$ to 210°C. The selenium layer is given a thin coat of vacuum-deposited cadmium, gallium or indium, on which, finally, a layer of metal is vacuum-deposited. When heated, the vacuum-deposited metal reacts with the selenium to produce an *N*-conduction layer of selenide about 50 μm thick on the surface of the selenium crystal. Thus, a *P-N* junction is formed at the interface between the selenide and *P*-type selenium. The thin layer of the vacuum-deposited metal is semitransparent and serves as a second electrode connected to a ring metal terminal.

In a silicon photocell, a layer of *P*-type silicon 0.4 to 1 μm thick is made by diffusing boron or aluminium into a wafer of *N*-type silicon 0.3 to 1 mm thick. At the interface between the layers, a *P-N* junction with a depletion layer about 0.05 μm thick is formed. The terminals of the *P*-type silicon are furnished by a vacuum-deposited titanium film which is then protected by a thin layer of silver. On the other side, a hole is etched for connection to the *N*-type silicon wafer.

Characteristics and Parameters. The current-illumination characteristic of a photocell is shown in Fig. 14-9*a*. According to Eq. (14-21) at $R_L = 0$, the $I_{ph} = f(\Phi)$ relation should be linear. With increasing luminous flux, however, it diverges from a straight line owing to the effect of the junction resistance, r_j, and the semiconductor bulk resistance, r_b, shown in the equivalent circuit of Fig. 14-10

where the generation of electron-hole pairs is represented by an equivalent current generator, I_{ph}.

According to Kirchhoff's current law, we may write for this circuit

$$\left.\begin{aligned} I_{ph} &= I_j + I \\ I_j/I &= (R_L + r_b)/r_j \end{aligned}\right\} \qquad (14\text{-}31)$$

Hence,

$$I = I_{ph} r_j / (R_L + r_j + r_b) \qquad (14\text{-}32)$$

At $R_L = 0$

$$I = I_{ph} / (1 + r_b/r_j) \qquad (14\text{-}33)$$

When Φ is low, r_j is much greater than r_b, so I = (approx.) I_{ph} = = (approx.) $s_{int}\Phi$, and the luminous characteristic is almost linear.

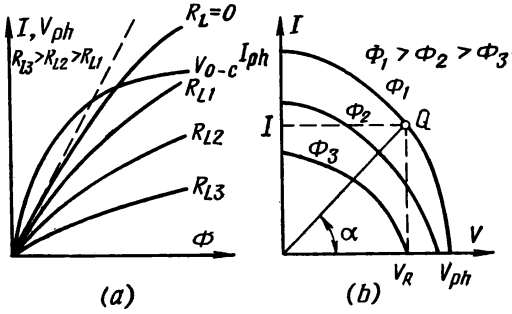


Fig. 14-9. (*a*) Current-illumination and (*b*) volt-ampere characteristics of a photocell

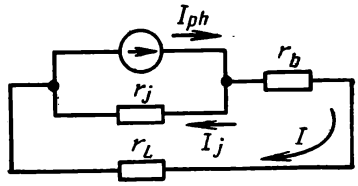


Fig. 14-10. Equivalent circuit for a photocell

As Φ increases, the junction resistance decreases and the $I = f(\Phi)$ relation diverges from a straight line still more.

Figure 14-9*a* also shows the $V_{ph} = \varphi(\Phi)$ characteristic plotted according to Eq. (14-28).

A family of *volt-ampere characteristics* of a photocell is shown in Fig. 14-9*b*. They represent a portion of the volt-ampere characteristics of an illuminated *P-N* junction (see Fig. 14-7). At a specified luminous flux, Φ_1, the characteristic makes on the current axis an intercept equal to the photocurrent $I_{ph} = I_{s\text{-}c}$, and on the voltage axis an intercept equal to the photo-emf $V_{ph} = V_{o\text{-}c}$.

Using the family of volt-ampere characteristics, we may construct a load line passing through the origin. The cotangent of the angle that the load line makes with the voltage axis, cot α, is proportional to the load resistance R_L. The point of intersection between the load line and the volt-ampere characteristic is the *Q*-point; it is located by the current in the external circuit, I, and the voltage across the load resistance, V_R. The area of the rectangle bounded by the coor-

dinate axes and the perpendiculars erected from the Q-point to the axes is proportional to the power dissipated in the external circuit.

Figure 14-11 shows the *spectral response curves* of basic industrial photocells together with a solar energy curve, a curve representing

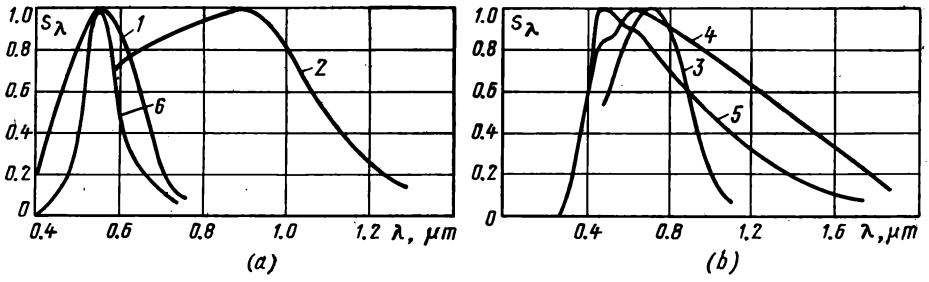


Fig. 14-11. Spectral response of photocells

1—**selenium**; *2*—silver sulphide; *3*—silicon; *4*—number of photons in solar flux; *5*—solar spectrum; *6*—response of human eye

the number of photons in the solar flux, and the response of the human eye.

Figure 14-12 shows the *frequency response curve* of a photocell. As is seen, an increase in the frequency, f, sinusoidally modulating the incident luminous flux in intensity brings about a decrease in the sensitivity of the photocell—a fact accounted for by the time lag mainly traceable to the recharging time constant of the barrier capacitance in the P-N junction. On the frequency response curve, we can mark the limiting frequency, f_{lim}, at which the sensitivity drops to $1/\sqrt{2}$ of the value at $f = 0$.

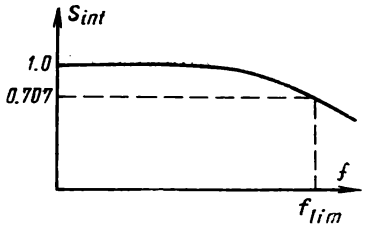


Fig. 14-12. Frequency response of a photocell

Photoelectric cells may be used as sources of electric energy or photoresistive devices.

In the latter case, the parameters of importance are their spectral characteristic, threshold flux and detectability (see Sec. 14-4).

For photocells used as sources of electric energy, the most important parameters are their volt-ampere characteristic and the efficiency of conversion of light energy into electric energy

$$\eta = IV_R/\Phi \qquad (14\text{-}34)$$

The efficiency of a photocell depends on a number of factors. The most pronounced effect is exerted by the loss of light, mainly by reflection, described in terms of the reflection coefficient. The reflected light energy does not contribute to conversion. The loss of light

also includes those photons which cannot generate electron-hole pairs as they are absorbed by the lattice, free charge carriers, excitons, etc.

Conversion is also accompanied by losses of energy, owing to recombination, generation of electron-hole pairs at a distance from the depletion layer exceeding the diffusion length, losses in the junction and bulk resistances, and so on.

The efficiency of a photocell increases with increasing luminous flux and photoelectromotive force. However, when Φ is strong, a rise in the concentration of free charge carriers raises the probability that they will recombine and also reduces the collection factor $\varkappa$. In addition, when a photocell is exposed to a strong Φ, its temperature increases, so that I_r builds up and reduces the efficiency.

The limit for the rise in the photo-emf and V_R is set by the potential barrier of the P-N junction. So, V_R may be raised if we use highly alloyed semiconductor materials having a wide forbidden band. Then, the Fermi level will approach the bottom of the conduction band in an N-type semiconductor and the top of the valence band in a P-type semiconductor, so that $e\varphi_c$ = (approx.) ΔE_f. Theoretically, at $N_a = N_d = 10^{19}$ cm^{-3}, the efficiency is about 25%.

The proper choice of a semiconductor material is especially important for solar-battery units which handle a very wide spectrum. In more detail, this will be discussed later in the section.

Parameters of Various Photocells. The most commonly used photocells are selenium and silver sulphide types. The construction of a selenium photocell is shown in Fig. 14-8*a*. These are used mostly in cine and photo equipment because their spectral response is similar to that of the human eye (Fig. 14-11*a*).

The integrated sensitivity of selenium photocells is anywhere between 300 and 700 μA/lm. The photo-emf of such devices does not exceed 0.5 or 0.6 V. The limiting frequency, f_{lim}, is about several hundred hertz.

In silver sulphide photocells, the P-N junction is made up of a semitransparent film of gold and a film of silver sulphide applied to a metal base plate. These photocells are very sensitive to long-wave radiation in the visible and infrared regions of the spectrum.

The integrated sensitivity of silver sulphide photocells ranges between 10 and 12 mA/lm. Like selenium devices, they have a low efficiency (1 to 2%), so they are not used as sources of electric energy. One of the reasons for such a low efficiency is the small diffusion length in polycrystalline semiconductor films.

Weak light signals are detected by photovoltaic cells based on single germanium and silicon crystals, and also semiconductor compounds, such as indium antimonide (InSb) and indium arsenide (InAs). Here, P-N junctions are produced by alloying (for example, by fusing indium into N-type germanium) or local impurity diffusion.

In order to boost the small electric signals produced in detecting weak luminous fluxes, it is desirable to have a high load resistance in the external circuit. The load resistance, however, is limited by the internal resistance of the photocell, R_{int}, which is determined by the resistance of the reverse-biased junction. When $R_L \ll R_{int}$, the energy characteristic is more linear (see Fig. 14-9) and the recharging time constant of the barrier capacitance decreases.

In this respect, silicon photocells are preferable because the forbidden band in silicon (about 1.1 eV) is approximately half as wide again as that in germanium, so the reverse current through the photocell is lower. Because of this, germanium photocells are used at temperature close to the temperature of liquid nitrogen (77 K).

Silicon devices are most sensitive to radiation at λ = (approx.) 0.8 μm; their longwave limit is λ = (approx.) 1.1 μm and the specific detectability is D^* = (approx.) 10^{13} W^{-1} cm $Hz^{1/2}$.

For operation in the infrared region of the spectrum use is made of photocells manufactured from semiconductor materials with a relatively narrow energy gap (InSb and InAs). For example, a photocell made of indium antimonide has a maximum sensitivity at λ = (approx.) 5.5 μm, and its specific detectability is about 5×10^{11} W^{-1} cm $Hz^{1/2}$.

Solar batteries use silicon units based on single silicon crystals, and also film units based on cadmium sulphide. The major requirements for solar-battery units are as follows.

First of all, their spectral response should be as close to the solar spectrum as possible. The spectral response of a semiconductor material markedly depends on the width of its forbidden band, ΔE_f. Referring to Fig. 14-11*b*, the spectral response of silicon, as well as that of cadmium sulphide, is sufficiently close to the solar spectrum.

Another important requirement is that the efficiency of solar-battery units should be as high as possible. As already noted, the efficiency depends on many factors. As is shown in reference [14], when R_L is approximately equal to R_{int}

$$\eta = s_{int}^2 \Phi R_{int}/4 \qquad (14\text{-}35)$$

As is seen, the sensitivity of a material used in a solar-battery unit should be maximal and its R_{int} should also be high. According to these conditions and the requirement for the spectral response, the optimal width of the forbidden band $\Delta E_{f,opt}$, ranges between 1.1 and 1.6 eV. From this point of view, the best performance is shown by gallium arsenide ($\Delta E_f \approx 1.4$ eV), silicon ($\Delta E_f \approx 1.1$ eV), and cadmium sulphide ($\Delta E_f \approx 2.0$ eV). As a rule, solar batteries use these materials.

The construction of a silicon photocell is shown in Fig. 14-8*b*. The efficiency of solar batteries based on silicon units reaches 15

to 19%, and that of batteries made of gallium arsenide runs into 13%.

Major disadvantages of solar-battery units made of single semiconductor crystals are a small operating surface area (a few square centimetres) and a low output power-to-mass ratio (about 50 W/kg).

Film solar-battery units based on cadmium sulphide have a higher power-to-mass ratio (about 200 W/kg) and a larger operating surface area, but a lower efficiency (about 8%).

14-7. PHOTODIODES

Definition. A *photodiode* is a semiconductor diode in which the reverse current increases whenever light falls on the diode.

Design. A likely construction of a photodiode and its connection in a circuit are shown in Fig. 14-13*a* and *b*. It consists essentially

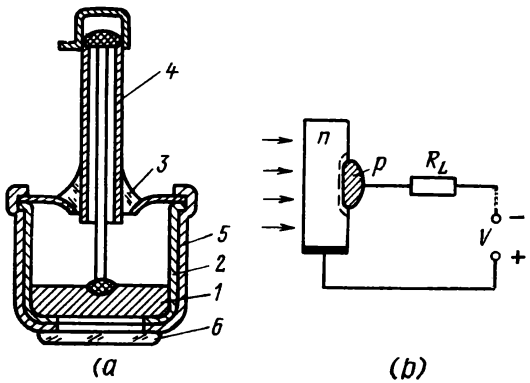


Fig. 14-13. Photodiode
(*a*) construction; (*b*) connection in a circuit; *1*—germanium wafer; *2*—support; *3*—insulator; *4*—bushing; *5*—metal case; *6*—glass window

o a round wafer of *N*-type germanium serving as the base enclosed in a metal case at a glass-covered window. The *P-N* junction is formed by fusing a dot of indium into a germanium wafer. In this type of photodiode, the luminous flux is at right angles to the *P-N* junction. Another design is available in which the luminous flux is parallel to the junction.

As the active element, photodiodes may use various junctions such as an abrupt symmetrical *P-N* junction, *P-I-N* junction, *P-N* junction with a nonuniformly doped emitter, metal-semiconductor (Schottky) junction, heterogeneous junction, and so on.

Principle of Operation. For its operation, a photodiode depends on the physical processes occurring in an illuminated junction (see Sec. 14-5). When no light strikes the diode ($\Phi = 0$), the reverse voltage gives rise to a reverse current whose value, as will be recalled, depends on the number density of minority carriers in the semiconductor regions, the junction area, the physical processes in the barrier region, etc. These were discussed in Sec. 11-3 when we

examined the volt-ampere characteristic of a real semiconductor diode. Since, in the absence of illumination, the photodiode does not practically differ from a conventional semiconductor diode, everything said in Sec. 11-3 holds for the $I = f(V)$ characteristic of the photodiode, provided that $\Phi = 0$. This relation is sometimes called the *dark volt-ampere characteristic.*

When the photodiode is illuminated ($\Phi > 0$), the photons striking its base (the *P*-type germanium wafer in the device shown in Fig. 14-13) generate electron-hole pairs, most intensively near the base surface. The newly formed electrons and holes diffuse through the bulk of the base towards the *P-N* junction. The holes are captured by the junction field, $\mathscr{E}_j$, and driven out to the *P*-region, thereby increasing the minority carrier flow density and, as a consequence, the reverse current in the photodiode. In order to enable most of the newly formed holes to reach the *P-N* junction, the base region should be narrower than the diffusion length of holes: $w < L_p$.

As has been shown in Sec. 14-5, this process is most effective when

$$w = \text{(approx.)}\ 1/\alpha_0$$

The thickness of the photodiode base is chosen in accordance with this requirement.

Volt-Ampere Characteristics. To define the relation between the current in a photodiode and the applied voltage, we shall use Eq. (14-25)

$$I = I_{ph} - I_r [\exp(eIR_L/kT) - 1]$$

The external circuit of Fig. 14-4*b* contains an additional voltage source and, as a consequence, the voltage across the *P-N* junction is

$$V_j = IR_L - V \qquad (14\text{-}36)$$

Using the above equation, we may re-write Eq. (14-25) as follows

$$I = I_{ph} - I_r \{\exp[e(IR_L - V)/kT] - 1\} \qquad (14\text{-}37)$$

The volt-ampere characteristics plotted according to Eq. (14-37) are shown in Fig. 14-14*a*. As is seen, they represent a portion of the volt-ampere characteristic of an illuminated *P-N* junction (Fig. 14-7).

As already noted, the dark volt-ampere characteristic ($\Phi = 0$) is the downward slope of the volt-ampere characteristic of the diode. So, in silicon photodiodes, the dark current is markedly smaller than in germanium devices, and a slight increase in this current with increasing reverse voltage is traceable to the same causes as in crystal diodes (see Sec. 11-3).

According to Eq. (14-27), the photocurrent rises linearly with rising luminous flux, so, when $\Phi > 0$ and $\Delta\Phi = \Phi_3 - \Phi_2 = \Phi_2 - \Phi_1 = \text{constant}$, the curves are spaced equal distances apart.

The *current-illumination characteristics* of a photodiode (Fig. 14-14*b*) are linear over a sufficiently wide range of luminous flux variations. As the reverse voltage rises, the barrier layer widens and the base shortens. As a result, a smaller proportion of minority carriers recombines in the bulk of the base as they diffuse towards the *P-N* junction, and the photocurrent increases.

The *relative spectral response curves* of a germanium and a silicon photodiode are shown in Fig. 14-14*c*. During illumination, electron-hole pairs are generated mainly because of intrinsic absorption (see

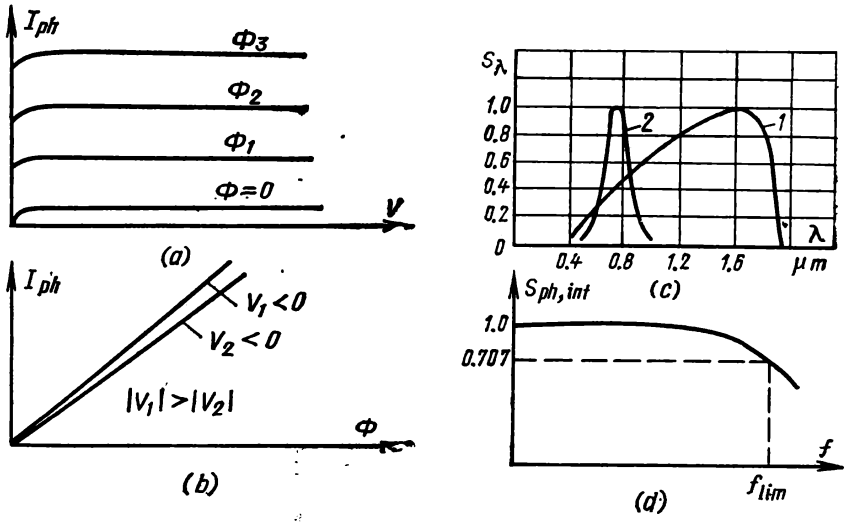


Fig. 14-14. (*a*) Volt-ampere, (*b*) current-illumination, (*c*) relative spectral and (*d*) frequency characteristics of photodiodes

1—germanium type; *2*—silicon type

Sec. 14-1). So, the spectral response of devices made of silicon whose forbidden band is wider ($\Delta E_f \approx 1.1$ eV) is a maximum at lower values of λ as compared with germanium photodiodes ($\Delta E_f \approx 0.72$eV). For the same reason, germanium devices have their longwave limit in the region of longer wavelengths.

At shorter wavelengths, the sensitivity of photodiodes somewhat drops because α_0 is sufficiently high ($\alpha_0 \approx 10^5$ cm^{-1}), photons are chiefly absorbed near the base surface where the probability that they will recombine at the surface centres is rather high, and the number of minority carriers arriving at the *P-N* junction is decreased.

The *frequency response curve* of a photodiode (Fig. 14-14*d*) illustrates its behaviour when it is illuminated by an intensity modulated luminous flux. The frequency, *f*, at which the flux is modulated is laid off as abscissa. A decrease in the photocurrent with

increasing frequency is a testimony to a time lag in the response of the device.

The time lag in the response of photodiodes is mainly traceable to the charge time constant of the junction capacitance, the time of diffusion of carriers towards the junction, t_D, and the time during which carriers drift through the space charge region in the junction, t_{dr}. If the period of modulation, T_m, is comparable with the total time of carrier motion, $t_{tot} = t_{dr} + t_D$, variations in the current will lag behind, as it were, variations in the luminous flux intensity. So, as the frequency rises, the amplitude of the alternating current component in the photodiode load will fall, and the phase shift between the modulating oscillations and the alternating component will increase. It is shown [23] that when t_{tot} = (approx.) $0.4\ T_m$, the current amplitude decreases to $1/\sqrt{2}$ of its value at a low modulation frequency and the phase shift exceeds 70°.

In our photodiode, the space charge region in the junction is sufficiently narrow, and the time of carrier motion is mostly determined by the time of hole diffusion in the base (N-type germanium), t_{tot} = (approx.) t_D. Since, in contrast to a conventional crystal diode, holes in our case diffuse from the base to the junction, t_D is twice the value determined according to Eq. (9-121). So, recalling Eq. (14-23), we may write

$$t_D = w^2/D_p = 1/\alpha_0^2 D_p \qquad (14\text{-}38)$$

Hence, the limiting frequency at which the amplitude of the alternating current component decreases to $1/\sqrt{2}$ ($t_{tot} \approx 0.4\ T_m$) is

$$f_{lim} = 1/t_{tot} = 0.4\,\alpha_0^2 D_p \qquad (14\text{-}39)$$

In semiconductors, the mobility of electrons is generally higher than that of holes and, as a consequence, $D_p < D_n$. So, as far as their transient recovery is concerned, photodiodes with an N-type semiconductor base are preferable.

Photodiode Parameters and Ratings. The ratings specified in data sheets for photodiodes include the nominal operating voltage, V_{op}, and the maximum safe reverse voltage, $V_{r,\,max}$, which guarantees breakdown-free operation of a photodiode.

Photodiodes use the same parameters as photoresistors, because they, too, form electrical signals when struck by light and detect light signals. Their parameters are the sensitivity, s, the limiting frequency, f_{lim}, the threshold flux, Φ_{thr}, and the detectability, D.

Recently, to improve the parameters, a number of new photodiodes differing from those discussed above have been designed.

The Photodiode with a Nonuniformly Doped Base. The energy diagram of this photodiode is shown in Fig. 14-15. The P-N junction is formed here by diffusing acceptor impurities in an N-type semicon-

ductor wafer. The impurity concentration in the base is nonuniform, decreasing from the surface to the junction. Due to the thermal ionization of acceptor atoms, the hole concentration in the base is also nonuniform. The holes diffuse towards the junction, thereby leaving behind the immobile negative charges of ionized acceptor atoms at the base surface. As a result, an electric field, $\mathscr{E}_b$, whose strength vector is directed towards the base surface is developed in the base. The charge carriers generated by the luminous flux incident on the base are separated by this field. Now, electrons not only diffuse to

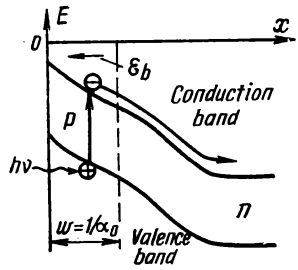


Fig. 14-15. Energy-band diagram of a photodiode with the nonuniformly doped base

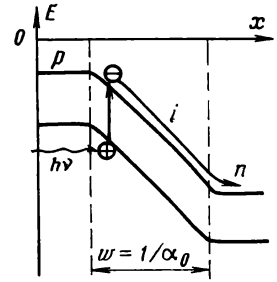


Fig. 14-16. Energy-band diagram of a *P-I-N* photodiode

the junction, but at the same time drift in $\mathscr{E}_b$. Accordingly, the total time of their motion in the base decreases, and f_{lim} increases.

These photodiodes are manufactured by diffusion, so the thickness of the base may be reduced to 3 or 5 μm which also improves their frequency properties.

The P-I-N Photodiode. This may be both germanium and silicon. The *P-I-N* photodiode differs from conventional types in that the *P*- and *N*-regions are separated by a layer of an intrinsic semiconductor. So, there are two junctions in this device, namely *P-I* and *N-I*. However, if the photodiode operates at a reverse voltage such that the space charge regions of both junctions, which mainly occupy the *I*-layer, overlap (Fig. 14-16), we speak of only one junction whose depletion layer lies in the *I*-region.

The base in such devices is made sufficiently thin, so that absorption mainly occurs in the *I*-region. The time of motion of the generated carriers is determined by the time of their drift through the *I*-region:

$$t_{tot} = (\text{approx.})\, t_{dr} = w/v_{dr} = 1/\alpha_0 v_{dr} \qquad (14\text{-}40)$$

whence we may write for $f_{lim} = 0.4/t_{tot}$

$$f_{lim} = (\text{approx.})\, 0.4\, \alpha_0 v_{dr} \qquad (14\text{-}41)$$

where $v_{dr} = \mu_n \mathscr{E}_c$ and $\mathscr{E}_c$ is the contact field in the junction.

Along with photodiodes utilizing symmetrical *P-N* junctions, *P-I-N* photodiodes are successfully used to detect signals with frequencies up to 10 GHz.

Among the advantages of *P-I-N* photodiodes, we may mention a higher peak-reverse voltage and a lower junction capacitance.

The Schottky-Barrier Photodiode. The construction and energy-band diagram of this diode are shown in Fig. 14-17. It consists essen-

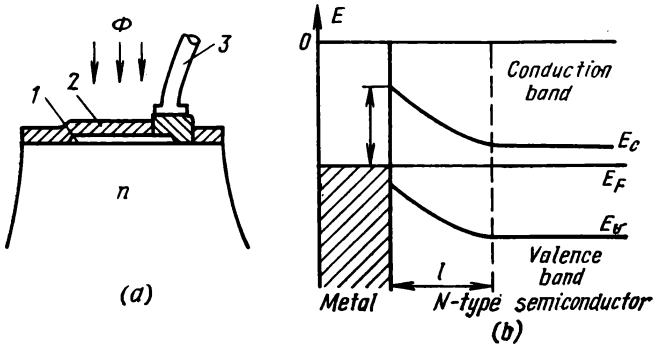


Fig. 14-17. (*a*) Construction and (*b*) energy-band diagram of a Schottky-barrier photodiode
1 — metal film; *2* — transparent coating; *3* — lead

tially of an *N*-type silicon crystal given a thin coat of gold (about 0.01 μm) which, in turn, is covered by a thin transparent dielectric (zinc sulphide) film (about 0.05 μm). Owing to the difference in refractive index between silicon, gold and zinc sulphide, a light beam having a certain wavelength is reflected by the interfaces between these films and penetrates through the metal film into the silicon crystal with almost none of its power lost. For example, the luminous flux emitted by a gelium-neon laser ($\lambda \approx 0.63$ μm) loses only 5% of its power.

If the energy of a photon exceeds the width of the forbidden band, $h\nu > \Delta E_f$, intrinsic absorption occurs in the silicon crystal at its surface. The generated electrons drift in the junction field, thereby constituting the photocurrent.

When a photodiode operates with a high reverse bias, the carriers moving in the barrier region may have enough energy to break valence bonds (impact ionization is said to take place). In this case, avalanche multiplication typical of the avalanche breakdown occurs in the diode (see Sec. 10-5). This phenomenon is utilized in avalanche photodiodes which may be based not only on the Schottky barrier, but also on a conventional *P-N* junction.

The Avalanche Photodiode. Figure 14-18 shows the construction of a planar silicon avalanche photodiode using a *P-N* junction and a platinum-silicon photodiode using a Schottky barrier. In both

diodes, the light-sensitive area is made as a small-diameter round window (40 to 60 μm). This area is made small so that avalanche multiplication could take place at a certain reverse voltage in the whole crystal volume near the illuminated surface. If the light-sensitive area were large, it would be difficult to obtain a uniform thin film (0.1 to 0.3 μm) of N^+-type silicon or PtSi at the crystal surface. In a nonuniform film, microregions with a low breakdown voltage might exist, avalanche multiplication would take place only in a small volume of the crystal, and the photocurrent density would markedly decrease.

To prevent photodiodes from local avalanche breakdown caused by decreased voltages at the film edges, use is made of a protective

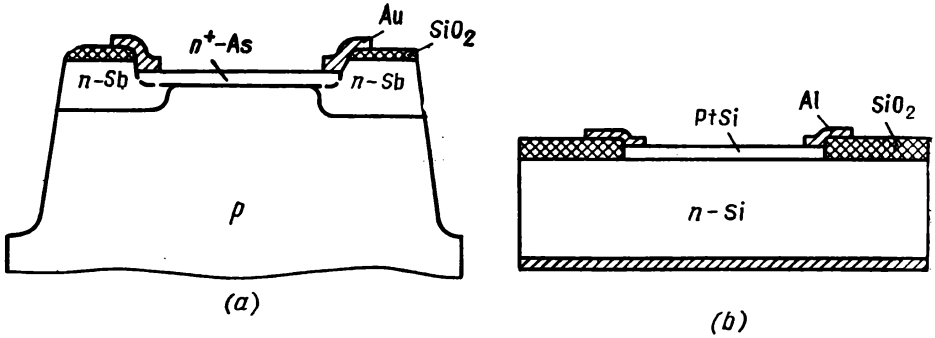


Fig. 14-18. (*a*) Construction of a silicon avalanche photodiode with a protective ring, and (*b*) construction of a platinum-silicon avalanche Schottky-barrier photodiode

ring, that is, a semiconductor ring having the same conduction as the main crystal but a lower impurity concentration. Owing to the smaller impurity concentration gradient in and a larger curvature radius of the protective ring, the breakdown occurs around its circumference at higher voltages than in the active area.

The physical processes in the avalanche photodiode differ from those in a conventional photodiode in the additional avalanche multiplication of the generated carriers in the depletion layer of the *P-N* junction. Because of this, the current, $I_{ph,\,aval}$, flowing in the external circuit of the photodiode is M times the current due to the generated electron-hole pairs and dark current, $I_{tot} = I_{ph} + I_d$, where

$$M = I_{ph,aval}/I_{tot} \qquad (14\text{-}42)$$

is the carrier multiplication factor.

In the literature of the subject, this is sometimes called the *primary photocurrent amplification* or *photomultiplication.*

Similarly to Eq. (10-59), M is defined as the ratio of the number of holes (electrons) generated by the avalanche to the number of primary holes (electrons) which caused the avalanche.

The polarity of the avalanche depends on the manner in which photons are absorbed in the photodiode. As is stated in Sec. 14-2, α_0 varies from wavelength to wavelength. Say, the absorptance of silicon at $\lambda =$ (approx.) 0.4 μm is about 5×10^4 cm^{-1}, whereas at $\lambda =$ = (approx.) 0.85 μm, it decreases to 7×10^2 cm^{-1}. As will be recalled, absorption takes place in a layer approximately $1/\alpha_0$ thick. So, in an N^+-P diode, the quanta of shortwave light ($\lambda \approx 0.4$ μm) are absorbed in the N^+-layer located near the surface. As a result, minority carriers—holes—move across the junction, thereby giving rise to a hole avalanche. In the case of longwave radiation, absorption mainly occurs past the junction and, as a consequence, electrons arrive at the space charge region to cause an electron avalanche.

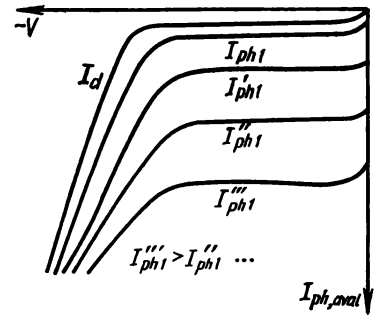


Fig. 14-19. Volt-ampere characteristics of an avalanche photodiode

The multiplication factor as a function of applied voltage is described by an expression similar to Eq. (10-60)

$$M = 1/[1 - (V_j/V_{bd,aval})^b] \qquad (14\text{-}43)$$

where $V_j = V - I_{ph,aval}R$ is the voltage across the junction, and R is the series bulk resistance of the photodiode. The values of b for various materials are given in Sec. 10-5.

Using the above expression, we may derive a formula for the maximum multiplication factor given in reference [23]

$$M_{max} = \text{(approx.)}\sqrt{V_{bd,aval}/nI_{tot}R} \qquad (14\text{-}44)$$

Since $I_{tot} = I_{ph} + I_d$, a high M can be obtained by reducing the dark current, I_d, and the bulk resistance, R. In avalanche photodiodes using the Schottky barrier, M ranges between 30 and 35.

Using Eqs. (14-42) and (14-43), assuming R to be low and, as a consequence, $V_j =$ (approx.) V, we may derive an expression describing the volt-ampere characteristics of the photodiode (Fig. 14-19)

$$I_{ph,aval} = I_{tot}/[1 - (V/V_{bd,aval})^b] \qquad (14\text{-}45)$$

14-8. PHOTOTRANSISTORS

Definition. The *phototransistor* is a semiconductor device having two P-N junctions, whose current increases owing to mobile charge carriers generated by incident light.

Design. A likely construction of the phototransistor is shown in Fig. 14-20*a*. Referring to the figure, the phototransistor differs from

a conventional transistor only in that it has a transparent window in its case. Incident light passes through this window and strikes the semiconductor wafer serving as the base. In the centre of this wafer, there is a collector junction formed by alloying. Other arrangements of phototransistor electrodes are also available. For example, a ring collector may be located at the illuminated base surface.

The connection of a phototransistor in a common-emitter circuit with the base open-circuited is shown in Fig. 14-20*b*. Here, the emitter junction is forward-biased and the collector junction is reverse-biased, so the phototransistor operates in the active region. In spite

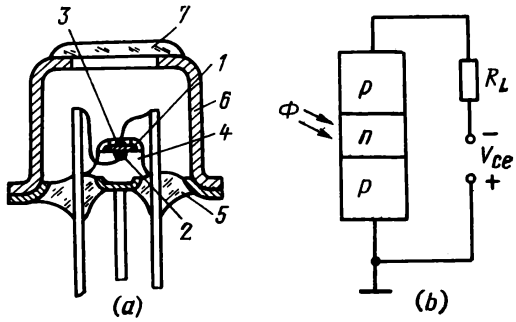


Fig. 14-20. Phototransistor

(*a*) construction; (*b*) connection in a circuit; *1*—*N*-type germanium wafer; *2*—emitter; *3*—collector; *4*—support; *5*—insulator; *6*—case; *7*—glass lens

of this, when no light strikes the transistor ($\Phi = 0$), the holes injected from the emitter are not entirely balanced by the electrons of the open-circuited base, and the current through it is low. The unbalanced holes space charge maintains the emitter junction potential barrier at a certain level, and a dark current, $I^e_{d,c}$*, flows in the transistor.

Principle of Operation. When the phototransistor is illuminated ($\Phi > 0$), electron-hole pairs are generated in the base owing to intrinsic absorption. Minority carriers—holes —diffuse to the collector junction and are swept away to the collector region where they contribute to the collector current in a manner similar to the photodiode. In contrast to the photodiode, however, the generated electrons which are the majority carriers in the base cannot leave the base because it is open-circuited. They accumulate in the base, thereby raising

* The phototransistor currents and voltages are designated as follows. The second letter in the subscript indicates the electrode whose voltage or current is being discussed. The upper letter indicates the common electrode. For example, V^e_b stands for the base voltage of a phototransistor connected in a common-emitter circuit, $I^b_{d,c}$ stands for the dark collector current in a phototransistor connected in a common-base circuit.

the negative space charge including that at the emitter junction. As a result, the potential barrier at this junction becomes lower, and holes begin to diffuse from the emitter into base. The holes diffuse into the base, move towards the collector junction and are swept by the collector junction field into the collector region.

Thus, the charge carriers generated in the illuminated base not only give rise to the photocurrent through the collector junction, but also spur the physical processes similar to those in a conventional transistor.

So, we may define the current in a phototransistor as

$$I^e_{tot,c} = h_{21b} I^e_{tot,e} + I^e_{d,c} + I^e_{ph}, \qquad (14\text{-}46)$$

When $I^e_{tot,b} = 0$, $I^e_{tot,e}$ is equal to $I^e_{tot,c}$, and Eq. (14-46) may be written as

$$I^e_{tot,c} = h_{21b} I^e_{tot,c} + I^e_{d,c} + I^e_{ph,c} \qquad (14\text{-}47)$$

whence

$$I^e_{tot,c} = I^e_{d,c}/(1 - h_{21b}) + I^e_{ph,c}/(1 - h_{21b}) \qquad (14\text{-}48)$$

As is seen, the current in the phototransistor is amplified by $1/(1 - h_{21b})$. When h_{21} = (approx.) 0.95 to 0.99 the value of $1/(1 - h_{21b})$ may run into several hundreds.

The connection of a phototransistor in a circuit (Fig. 14-20*b*) does not practically differ from that of a photodiode shown in Fig. 14-13*b*. If, however, we equip the phototransistor base with a lead, we shall be able to control the current in the transistor by varying the base current. In particular, this may be utilized for temperature compensation of the phototransistor parameters.

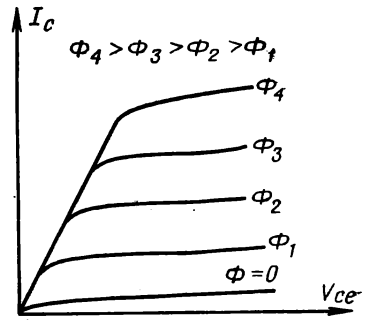


Fig. 14-21. Volt-ampere characteristics of a phototransistor

The *volt-ampere characteristics* of the phototransistor shown in Fig. 14-21 resemble the output characteristics of a conventional transistor connected in a common-emitter circuit, except that they are plotted for different fixed values of Φ, rather than I_b. The sharply rising portion of the characteristics represents the saturation region: when V_{ce} is low, the collector junction is rendered conducting by the holes accumulating in the collector as in a bipolar transistor. As in a bipolar transistor, the gradual slope of the characteristics is due to the base width modulation.

The *current-illumination characteristics* of the phototransistor, like those of the photodiode, are linear. As V_{ce} increases, the photocurrent somewhat rises owing to the base width modulation.

The *spectral response curves* are similar to those of the photodiode (see Fig. 14-14*c*).

The frequency properties of phototransistors are mainly determined by the diffusion of carriers in the base and the charging of the junction capacitances. As in photodiodes, the photocurrent decreases with increasing luminous flux modulation frequency.

Parameters. Like photoresistors and photodiodes, phototransistors are used to detect light signals. So, their operation as photodetectors is described in terms of the same parameters as in the case of photoresistors (see Sec. 14-4), namely threshold flux, Φ_{thr}, detectability, D, and others.

One of the most important parameters of the phototransistor is its *photocurrent gain* defined as the ratio of the collector photocurrent with the base open-circuited to the illuminated *P-N* junction photocurrent measured in the diode operation.

According to Eq. (14-48)

$$K_{ph} = 1/(1 - h_{21b}) \tag{14-49}$$

Another parameter of importance is the *current sensitivity* of the phototransistor, defined as the ratio of an increment in the output current to an increment in the incident flux with the input open-circuited and the output short-circuited for alternating current.

For the common-emitter arrangement, the current sensitivity is

$$h^e_{23} = \frac{dI^e_{tot,c}}{d\Phi}\bigg|_{I^e_{tot,b}=0} \tag{14-50}$$

The *voltage sensitivity* of the phototransistor, h^e_{13}, is defined as the ratio of an increment in the input voltage to an increment in the incident flux, determined under the same conditions.

14-9. LIGHT-EMITTING SEMICONDUCTOR DEVICES

Emission of Light by Semiconductors. For their operation light-emitting semiconductor devices depend on the emission of quanta of electromagnetic energy by charge carriers as they move from a higher to a lower energy state (*downward transition*). When discussing recombination of carriers in Sec. 9-3, we examined various types of downward transition (Fig. 9-9*a*), such as band-to-band transitions, transitions from the conduction band to an acceptor level or from a donor level into the valence band, transitions through deeply lying impurity levels (trap transitions), and so on. All these transitions may be accompanied by emission of energy in the form of either photons or phonons. For example, in accordance with the energy-band structure of a semiconductor (see Fig. 9-9*b* and *c*), both direct and indirect interband transitions may occur. In the case of direct downward transitions, the emission of photons is most likely to take place

(this is known as *emissive recombination*). With indirect transitions, the energy is generally released as phonons and is absorbed by the crystal lattice. As this takes place, the crystal temperature rises above the ambient temperature, and an equilibrium thermal emission takes place. Also, indirect transitions may result in the emission of photons in addition to phonons.

Transitions at which quanta of radiant energy are emitted are called *radiative transitions.* Radiative transitions are associated with *luminescence* and *stimulated emission.* The latter underlies the operation of quantum devices and is discussed in more detail in a course on microwave electronic circuits and quantum devices.

Luminescence is widely used in incoherent light-emitting semiconductor devices.

Luminescence is the emission of light in excess of thermal emission at a given temperature of a material with a duration considerably longer than the period of light oscillations.

For luminescence to take place, a semiconductor crystal should be excited by an external energy source, so that its internal energy would exceed the equilibrium value corresponding to a given temperature. In particular, excitation of a semiconductor is accompanied by the generation of non-equilibrium carrier concentrations, namely electrons in the conduction band and holes in the valence band.

In contrast to stimulated emission whose duration approaches the period of light waves (about 10^{-15} s), luminescence is associated with a rather long phosphorescence which is the emission of light from a substance after excitation has been removed. This is because in luminescence, the absorption of energy quanta is separated in time from the emission by intermediate processes. In addition, luminescence is accompanied by the emission of incoherent light with a relatively wide spectrum (about 10^{-2} μm), whereas the stimulated emission of an optical quantum oscillator (laser) is coherent and has a considerably narrower spectrum (less than 10^{-5} μm).

A semiconductor crystal may be excited by several mechanisms. If, for example, a semiconductor is excited due to the absorption of photons, luminescence is called *photoluminescence.* When a crystal is excited by a steam of fast electrons or other particles, we speak of *cathodoluminescence* or *radioluminescence.* There also is chemi- and bioluminescence.

In semiconductor electronics, of particular interest is *electroluminescence* where a crystal is excited by an electric field or current. For example, an electric field may cause band-to-band excitation in which the main mechanisms are impact ionization by accelerated electrons and also the release of electrons from trapping centres.

A semiconductor crystal containing a *P-N* transition region may be excited by injection of minority carriers under the action of an

external forward voltage source, avalanche breakdown of the *P-N* transition region, Zener breakdown, etc.

The spectrum of luminescent emission depends on the nature of the predominant radiative transitions. Band-to-band transitions are characterized by the spontaneous emission spectrum which has its peak at the frequency $\nu \geqslant \Delta E_f/h$. Most probable are radiative transitions involving local impurity energy levels or lattice imperfections. Such transitions may take place, say, due to the recombination of an electron captured by an impurity level at the bottom of the conduction band with a hole in the valence band, or due to the recombination of a hole at a local level near the bottom of the valence band with an electron from the conduction band.

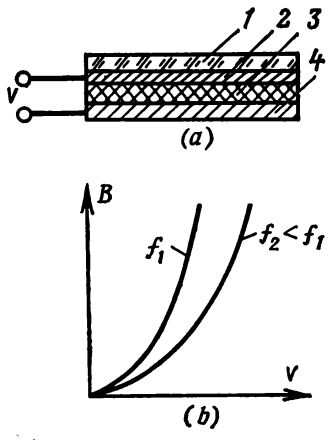


Fig. 14-22. (*a*). Construction and (*b*) voltage-brightness characteristic of a luminescent capacitor

1—glass base layer; *2*—semitransparent metal film; *3*—phosphor; *4*—metal film

The peaks of emission spectra in transitions involving local levels are shifted into the longwave region relative to the peak of spontaneous emission.

In most cases, radiative transitions in semiconductors are accompanied by nonradiative transitions. Because of this, not all of the energy expended to excite a semiconductor is converted into luminous energy. The efficiency of luminescence is determined by the ratio of the output luminous energy to the total excitation energy. It increases with increasing number of local levels taking part in radiative transitions, and decreasing distance between these levels and the boundaries of the respective bands, because, in the latter case, electrons and holes can be captured more easily. The efficiency of luminescence also depends on the energy gap in a radiative transition; it should be much wider than kT.

The Light-Emitting Capacitor. The construction of this device used for visual indication of electric signals is shown in Fig. 14-22*a*. It consists essentially of a layer of phosphor (luminescent material capable of emitting light under the action of an electric field) sandwiched between a semitransparent metal film and a thin metal plate serving as the terminals. As a rule, the phosphor is zinc sulphide (ZnS) with an addition of copper and manganese. The terminals are connected to a source of an audio-frequency alternating voltage, v. The amplitude of this voltage should be large enough to develop a field in the phosphor, whose average strength would be 10^6 to 10^7 V/m. At some spots of the phosphor where the field may reach 10^8 V/m at the voltage peaks, the impact and field ionization of trapping cen-

tres begins. When v and, as a consequence, the field in the phosphor decrease, electrons move down, towards the levels of ionized centres. Some of these transitions are radiative, so the phosphor begins to glow.

The efficiency of luminescence in such devices is low, usually being 1 to 2%.

The main characteristic of electroluminescent (light-emitting) capacitors is the voltage-brightness characteristic showing the brightness, B, as a function of the applied voltage at various frequencies (Fig. 14-22*b*).

Using different combinations of additions to zinc sulphide, we may get light-emitting capacitors with yellow, red, blue or green luminescence.

The Light-Emitting Diode. This is a light-emitting semiconductor device having one *P-N* junction, intended to convert electric energy directly into the energy of incoherent radiation.

As a rule, light-emitting diodes are manufactured from silicon carbide (SiC), gallium phosphide (GaP) and gallium arsenide (GaAs). The emission spectrum depends on the width of the forbidden band of the material, and also on the type and concentration of impurities. Since in gallium arsenide direct interband transitions are predominant, in accordance with the width of its forbidden band ($\Delta E_f \approx$ $\approx$ 1.4 eV), its spectral response at 300 K has its maximum in the infrared region of the spectrum ($\lambda \approx 0.9$ μm).

Light-emitting diodes fabricated from silicon carbide emit yellow light ($\lambda \approx 0.6$ μm). The spectral response of the diode made of gallium phosphide with an addition of cadmium has several peaks. The compounds of cadmium and oxygen atoms form exciton levels (see Sec. 14-2) lying at the bottom of the conduction band. The radiative transitions between these levels and the acceptor levels of cadmium are accompanied by red luminescence ($\lambda \approx 0.7$ μm), and those between the donor and acceptor levels, by green luminescence ($\lambda \approx$ $\approx$0.55 μm). The spectral response of this material has its maximum corresponding to band-to-band transitions in the infrared region of the spectrum ($\lambda \approx 0.9$ μm).

P-N junctions in the semiconductor crystals for light-emitting diodes are fabricated by diffusion of impurities into a single semiconductor crystal or by alloying.

A likely construction of a light-emitting diode is shown in Fig. 14-23*a*. It consists essentially of a single semiconductor crystal having a *P-N* junction, located in the base of a lens forming a beam with a width of $\pm 20°$.

The semiconductor crystal is excited by the injection of minority carriers under the action of a forward voltage source connected to the junction.

A light-emitting diode may also have several luminescent crystals which are used in numerical readout indicators (Fig. 14-23*b*).

Light-emitting diodes have a high stability, low power consumption, long service life and miniature size. Owing to this, they are widely used as light indicators in microelectronics and also opto-

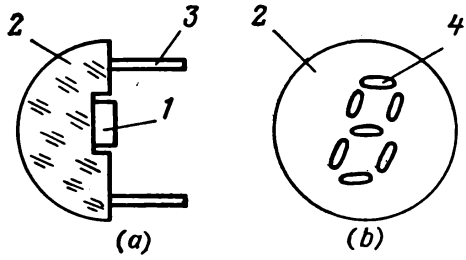


Fig. 14-23. (*a*) Construction of a light-emitting diode; (*b*) numerical readout light-emitting diode

electronics—a branch of radioelectronics dealing with electrical and optical methods of data processing. Optoelectronic methods and devices are studied in a course on microelectronics.

Chapter Fifteen

MISCELLANEOUS SEMICONDUCTOR DEVICES

15-1. NEGATIVE RESISTANCE DEVICES

Definition. A negative resistance device is that which has a region of negative slope in its volt-ampere characteristic. This negative slope may be due to the current falling with rising voltage (Fig. 15-1) or the voltage falling with rising current (Fig. 15-2). In the former

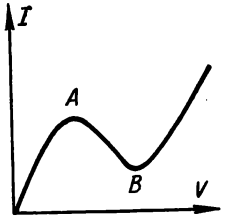


Fig. 15-1. Volt-ampere characteristic of a negative-resistance device (*N*-type)

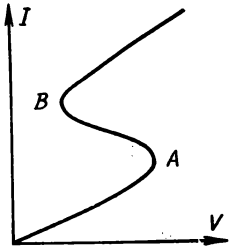


Fig. 15-2. Volt-ampere characteristic of a negative-resistance device (*S*-type)

case, we have what are known as voltage-controlled devices because their current is a univalued function of voltage, whereas the inverse relation, $V = f(I)$, is not. Sometimes such devices are referred to as *N-type* or devices with an *N-type response curve*. In the latter case, we have current-controlled devices, because their voltage is uniquely determined by the current, whereas the relation $I = \varphi(V)$ is not univalued. Such devices are sometimes called *S-type* or devices with an *S-type response curve*.

The class of negative-resistance semiconductor devices includes those with one, two or more *P-N* junctions, and also some devices operating by the bulk effects. Some of them (the *unijunction transistor*, the *thyristor* and the *tunnel diode*) will be taken up in this chapt-

er; others (the *avalanche diode* and the *Gunn diode*) mainly used at microwave frequencies are examined in a course on microwave and quantum electronic devices.

15-2. THE UNIJUNCTION TRANSISTOR

Definition. A unijunction transistor is a semiconductor device having only one *P-N* junction and two base leads (because of which it is also known as a double-base diode); it is intended for use as a switch.

Construction. A unijunction planar transistor manufactured by local diffusion is shown in Fig. 15-3*a*. The two bases of the device are in the bulk of an *N*-type silicon bar. Connections to the bases are

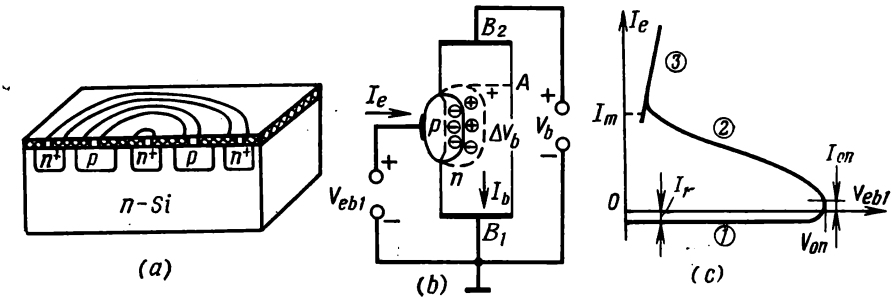


Fig. 15-3. Construction of a unijunction transistor (*a*); its equivalent circuit (*b*) and its volt-ampere characteristic (*c*).

by means of arc-shaped regions of N^+-type silicon and metallic ohmic contacts. The emitter is an arc-shaped region of *P*-type silicon in between. The P-N junction is formed at the interface between the *P*-type silicon region and the *N*-type region.

Consider the operation of this device by reference to its equivalent circuit in Fig. 15-3*b*, assuming that the silicon bar is cut across at the diameter of the ring-shaped electrodes.

Application of the base voltage, V_b, causes electrons to move towards the terminal B_2; as a result, a current, I_b, begins to flow in the base, producing a voltage drop across the base bulk resistance. Suppose that the voltage drop across part of the base between plane A and terminal B_1 is V_b. Applied in positive polarity to the *P-N* junction, this voltage drop biases the junction in the reverse direction. If V_{eb1} is less than ΔV_b, the junction is reversed-biased, and the current flowing through it, $-I_e = I_r$, is constituted by the minority carriers. As V_{eb1} increases, $-I_e$ remains unchanged so long as $V_{eb1} < \Delta V_b$ (portion *1* in the volt-ampere characteristic, Fig. 15-3*c*). When V_{eb1} exceeds ΔV_b, the junction is forward-biased, and the current flowing through it, I_e, is opposite to I_r. Owing to the injec-

tion of minority carriers into the base, the base bulk resistance decreases, and so does I_e. This entails a further decrease in ΔV_b, and so on. The rise of I_e is accompanied by a fall of the voltage between the emitter and terminal B_1, because the base bulk resistance between plane A and terminal B_1 is reduced. This is represented by portion *2* in the volt-ampere characteristic (see Fig. 15-3*c*); within this portion the device shows a negative incremental resistance. The subsequent rise in I_e corresponds to the volt-ampere characteristic of the *P-N* junction when it is forward-biased (portion *3*).

Should V_b is raised, the volt-ampere characteristic would shift to the right, because I_b and ΔV_b would also rise, and the junction would be forward-biased at a higher value of V_{eb1}.

Portion *3* of the volt-ampere characteristic represents the conducting ("on") state of the device: the voltage across it is low and an appreciable current is flowing through. The minimum current required to keep the device conducting is called the *maintaining current*, I_m; it is an important parameter of a conducting unijunction transistor.

Portion *1* of the volt-ampere characteristic represents the nonconducting ("off") state of the device: the current through the device is small, whereas the voltage across its electrodes is considerable.

A unijunction transistor will switch, that is, change over from the "off" to the "on" state, when $V_{eb1} = V_{on}$. The corresponding current is called the *turn-on current*, I_{on}, and is likewise a parameter of the device. At the point on the volt-ampere characteristic where $V_{eb1} = V_{on}$ the incremental (or dynamic) resistance of the device is zero, $r_d = dV_{eb1}/dI_e = 0$. Within portion *2*, the incremental resistance of the device is negative.

15-3. THYRISTORS

Definition. A thyristor is a bistable semiconductor device comprising three or more *P-N* junctions. It can switch between two states (reverse- and forward-voltage polarity) and its volt-ampere characteristic shows a region of negative resistance.

A thyristor having two leads is called a *dinistor*, and a thyristor with three leads is called a *trinistor*.

Construction. A likely thyristor design is shown in Fig. 15-4, whereas its diagram symbol, impurity distribution in the various regions of a device manufactured by alloying and diffusion, and the energy diagram of a state of equilibrium are shown in Fig. 15-5*a*, *b* and *c*.

The outer region where the current from the external circuit enters is called the *anode*; the opposite outer region is called the *cathode*. The intervening regions are usually called the *bases*. In a transistor, one of the bases is utilized to control the device and is called the *gate*.

Let us designate the P-N junctions in a thyristor as J_1, J_2 and J_3, and examine the physical processes that take place there (see Fig. 15-5*a*).

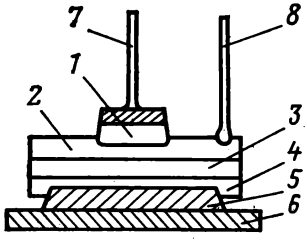


Fig. 15-4. Conctruction of a trinistor

1—P_1 egion; *2*—N_1 region; *3*—P_2 region; *4*—N_2 region; *5*—crystal holder; *6*—metal cathode lead; *7*—anode lead; *8*—gate lead

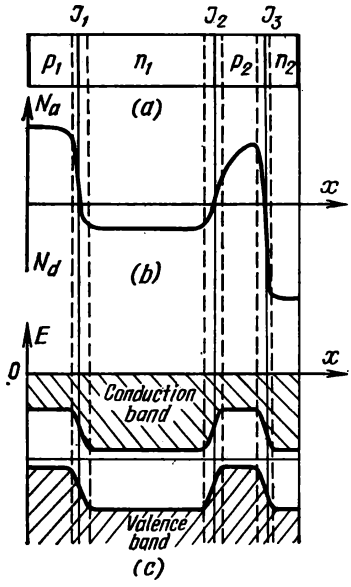


Fig. 15-5. Thyristor in sketch form (*a*); impurity distribution in a thyristor (*b*); energy-level diagram at equilibrium (*c*)

Operating Principle of a Dinistor. Suppose that a battery connected as shown in Fig. 15-6 applies a small voltage, V, between the anode and cathode of a dinistor. In the circumstances, J_1 and J_3 will be forward-biased and J_2 reverse-biased. Now that the state of equilibrium is disturbed, holes are being injected from region P_1 into region N_1 in junction J_1. In diffusing, the holes approach junction J_2, and the field $\mathscr{E}_k$ existing there sweeps them across the junction into region P_2. Thus, the P_1-N_1-P_2 structure is similar to that of a transistor, where the motion of charge carriers (holes from left to right in the diagram) may be described in terms of the current gain, α_1.

At the same time, the state of equilibrium existing in junction J_3 is also upset with the result that electrons are injected from region N_2 into region P_2 and are subsequently collected by region N_1. Thus, the N_2-P_2-N_1 structure may likewise be approximated by an equivalent transistor (in which electrons move from right to left in the diagram), with a current gain α_2. The holes in the P_1-N_1-P_2 transistor and the electrons in the N_2-P_2-N_1 transistor move in opposite directions. Because of this, the hole and electron components of the current flowing from anode to cathode, I_a, are in the same direction.

The further movement of holes from region P_2 to region N_2 is limited by the low potential barrier at junction J_3, whereas the motion of electrons from region N_1 to region P_1 is limited by the barrier at junction J_1. As a result, the base P_2 accumulates holes and the base N_1, electrons. So long as V is low, however, their space charges are not large enough to reduce the potential barrier at junction J_2. The latter remain reverse-biased, and the value of I_a is limited and equal to I_r.

Thus, as V increases within certain limits, the current through the device remains nearly unchanged (portion *1* in the volt-ampere characteristic of the dinistor in Fig. 15-7).

As V keeps rising, the processes taking place in the device change: the width of the barrier layer l at junction J_2 increases and the rise in V to some certain value creates conditions favourable for the ava-

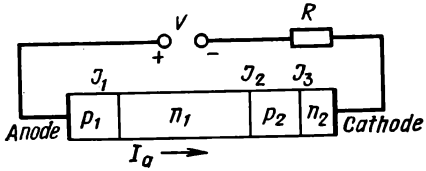


Fig. 15-6. Equivalent circuit of a dinistor

lanche multiplication of carriers in the barrier layer of J_2. The newly formed holes are swept by the field at J_2 across into region P_2, and electrons are swept across into region N_1. The current through the device builds up, and the N_1 and P_2 regions acquire excess concentrations of electrons and holes, respectively. As these concentrations rise, the potential barriers at J_1 and J_3 are reduced and so is the potential barrier at J_2. Hole injection across J_1 and electron injection across J_3 increase still more cumulatively, so that J_2 is rendered conducting. The increase in current through the device is accompanied by a fall in the resistance of the barrier layer at J_2 and also in the resistance of all the other regions of the device. Therefore an increase in the current through the device is accompanied by a fall in the anode-to-cathode voltage. In the volt-ampere characteristic (see Fig. 15-7) this is represented by portion *2* which is indicative of a negative resistance.

When junction J_2 is rendered conducting, the volt-ampere characteristic of the device is the same as the forward slope of the characteristic of a forward-biased diode (portion *3* in the plot of Fig. 15-7).

To establish the dependence of the current through the device on V, the dinistor may, as already noted, be approximated by a combination of two transistors (Fig. 15-8).

The currents in the P_1-N_1-P_2 transistor are connected by a relation of the form

$$I_{b1} = (1 - \alpha_1) I_a - I_{c01} \qquad (15\text{-}1)$$

where I_{c01} is the reverse (leakage or cut-off) collector current and α_1 is the current gain of this transistor.

For the N_2-P_2-N_1 transistor, we may write

$$I_{c2} = \alpha_2 I_c + I_{c02} \qquad (15\text{-}2)$$

where I_{c02} is the reverse (leakage or cut-off) collector current and α_2 is the current gain of this transistor.

Referring to Fig. 15-8, it is seen that $I_{b1} = I_{c2}$. Recalling that $I_a = I_c$ and designating the current through the device as $I = I_a = I_c$, we get

$$I = I_{c0}/(1-\alpha) \tag{15-3}$$

Here $I_{c0} = I_{c01} + I_{c02}$ is the reverse (leakage or cut-off) current through junction J_2 equal to the sum of the thermal current, thermal

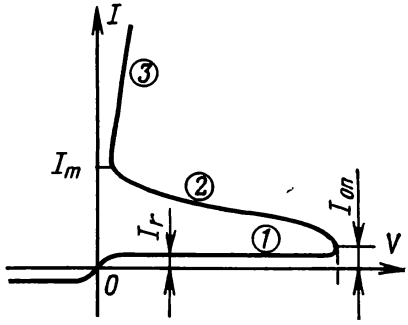


Fig. 15-7. Volt-ampere characteristic of a dinistor

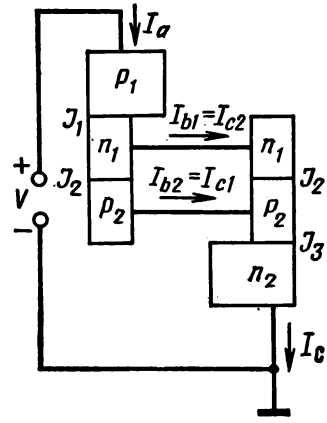


Fig. 15-8. Dinistor as two equivalent transistors

oscillation current and leakage current, and $\alpha = \alpha_1 + \alpha_2$ is the total current gain from junctions J_1 and J_3 to junction J_2.

Because the carriers experience avalanche multiplication at junction J_2, Eq. (15-3) may be re-written

$$I = MI_{c0}/(1-M\alpha) \tag{15-4}$$

Here M is the multiplication factor for electrons and holes at J_2, as defined by Eq. (10-59).

Using Eq. (10-60), we may re-write Eq. (15-4) thus

$$V = V_{av.b}\sqrt[b]{1-\frac{\alpha I + I_{c0}}{I}} \tag{15-5}$$

Equations (15-4) and (15-5) approximately describe the volt-ampere characteristic of a dinistor and may be used for an approximate evaluation of current and voltages within portions *1* and *2* of the characteristic.

Operating Principle of a Trinistor. As already noted, one of the middle regions in a thyristor is used to control the operation of the device; it is called the gate. The gate is ordinarily the middle region

of least width for which α is close to unity. In our example, this is region P_2 (see Fig. 15-5*a*).

Connection of a trinistor in a circuit is shown in Fig. 15-9. As is seen, the gate current I_g entering region P_2 is combined with the total current, which is equivalent to an increase in the current gain, α_2.

Equation (15-4) for this case may be written thus

$$I = \frac{MI_{c0} + M\alpha_2 I_g}{1 - M\alpha} \tag{15-6}$$

Similarly, Eq. (15-5) takes the form

$$V = V_{av.b} \sqrt[b]{1 - \frac{\alpha I + I_{c0} + \alpha_2 I_g}{I}} \tag{15-7}$$

It is seen from the above equations that V_{on} associated with an avalanche rise in current decreases with increasing I_g, whereas I

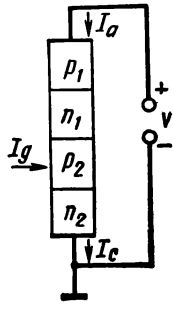


Fig. 15-9. Equivalent circuit of a trinistor

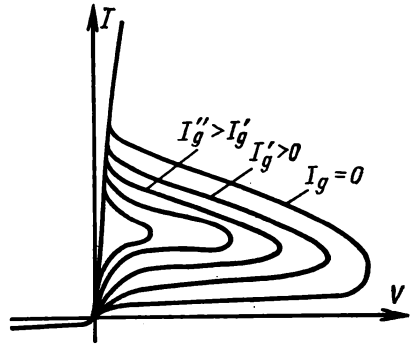


Fig. 15-10. Volt-ampere characteristics of a trinistor

at the point of connection of the thyristor increases ($r_d = dV/dI = 0$). Thus, by varying I_g, we can readily switch the trinistor between an "off" and an "on" state.

The volt-ampere characteristics of a trinistor are shown in Fig. 15-10.

Parameters and Ratings of Thyristors. The key parameters of thyristors are the voltages and currents corresponding to the characteristic points on their volt-ampere curves (see Figs 15-7 and 15-10). These are the turn-on voltage, V_{on}, the basic voltage at the turn-on point ($r_d = dV/dI = 0$); the turn-on current, I_{on}, the basic current at the turn-on point; the maintaining current, I_m, the minimum basic current required to hold a thyristor in the "on" state. The term "basic voltage" refers to that between the basic electrodes, that is the electrodes connected to load, whereas the term "basic current" refers to the current flowing through the basic electrodes.

Other important parameters are the time intervals describing the operation of a thyristor as a switch. These are the delay time t_d, the rise time t_r, the lag time t_l, the fall (or decay) time t_f, the gate turn-on time $t_{g,on}$, and the gate turn-off time $t_{g,\,off}$ can readily be understood from reference to Fig. 15-11 where the upper curves re-

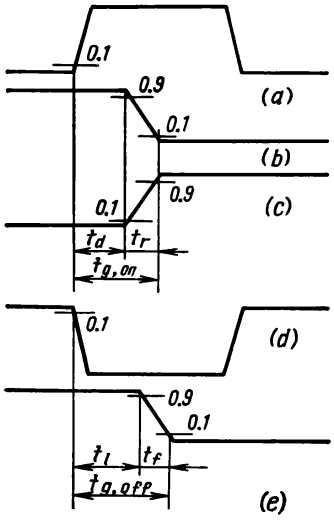


Fig. 15-11. Defining the switching parameters of a thyristor

(*a*) and (*d*)gate turn-on and turn-off pulses, respectively; (*b*) basic voltage amplitude; (*c*) and (*e*) basic current amplitude

present pulses at the gate and the lower curves are the waveforms of the basic current and voltage.

Commercially available thyristors come for voltages from a few volts (low-voltage units) to several kilovolts (high-voltage units). Accordingly, their parameters vary between wide limits, depending on thyristor application.

15-4. THE TUNNEL DIODE

Definition. A *tunnel diode* is a device based on a degenerate semiconductor in which quantum mechanical tunnelling results in a region of negative resistance in their volt-ampere characteristics with forward biasing.

Construction. Tunnel diodes are ordinarily manufactured by alloying from germanium or gallium arsenide. Radio-frequency and microwave tunnel diodes are built into cases designed for connection to resonant cavities, coaxial and waveguide transmission lines.

The source materials are highly doped semiconductor crystals with an impurity concentration of $N \approx 10^{19}$ cm^{-3} and more.

The energy-level diagram of a tunnel diode in the absence of an applied voltage is shown in Fig. 15-12. As will be recalled (see Sec. 9-2), impurity energy bands are formed in semiconductors with a high concentration of impurities. In an N-type semiconductor such

a band overlaps the conduction band; in a P-type semiconductor it overlaps with the valence band (Fig. 9-5). Because of this, the Fermi level in N-type semiconductors with a high concentration of impurities lies above E_c, whereas in P-type semiconductors it lies below E_v. As a result, in the interval $\Delta E = E_v - E_c$ (Fig. 15-12), any energy level in the conduction band of the N-type semiconductor corresponds to a similar energy level beyond the barrier layer, that is, in the valence band of the P-type semiconductor.

Owing to the high impurity concentration, the resulting P-N junction is very narrow ($l \leqslant 0.02$ μm). Thus, the particles in the

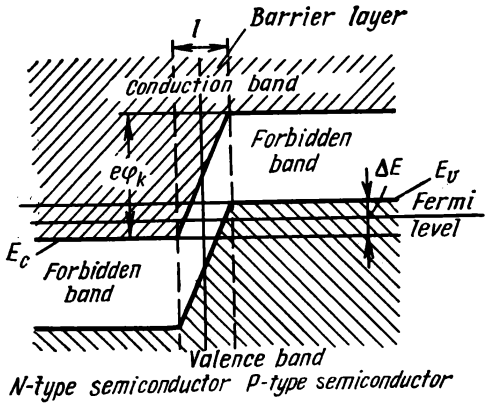


Fig. 15-12. P-N junction of highly doped semiconductors at equilibrium

N- and P-type semiconductors with energy states in the interval ΔE are separated by a narrow (l) and high (φ_c) potential barrier. In the valence band of the P-type semiconductor and in the conduction band of the N-type semiconductor some of the energy states in the interval ΔE are vacant. As a consequence, particles can travel across such a narrow potential barrier, with free energy states on either side, by quantum mechanical tunnelling. Obviously, on approaching the barrier, the particles experience repulsion and fall back in most cases, yet the probability of finding some particles across the barrier is nonzero, and the tunnel current density is $j_t \neq 0$.

Currents in a Tunnel Diode. At equilibrium (in the absence of an applied voltage), the total current across the junction is zero. However, instead of Eq. (10-5) describing the state of equilibrium for a conventional P-N junction,

$$j_D + j_d = 0$$

we must write

$$j_D + j_d + j_t = 0 \qquad (15\text{-}8)$$

Examine the quantum mechanical tunnelling of electrons separately, while bearing in mind that the results can fully be applied to the tunnelling of holes as well.

Figure 15-13*a* shows the energy-level diagram of a tunnel diode; on the right and left of the diagram are the energy state distributions, $S_q(E)$, shown by dashed lines. The areas bounded by the dashed

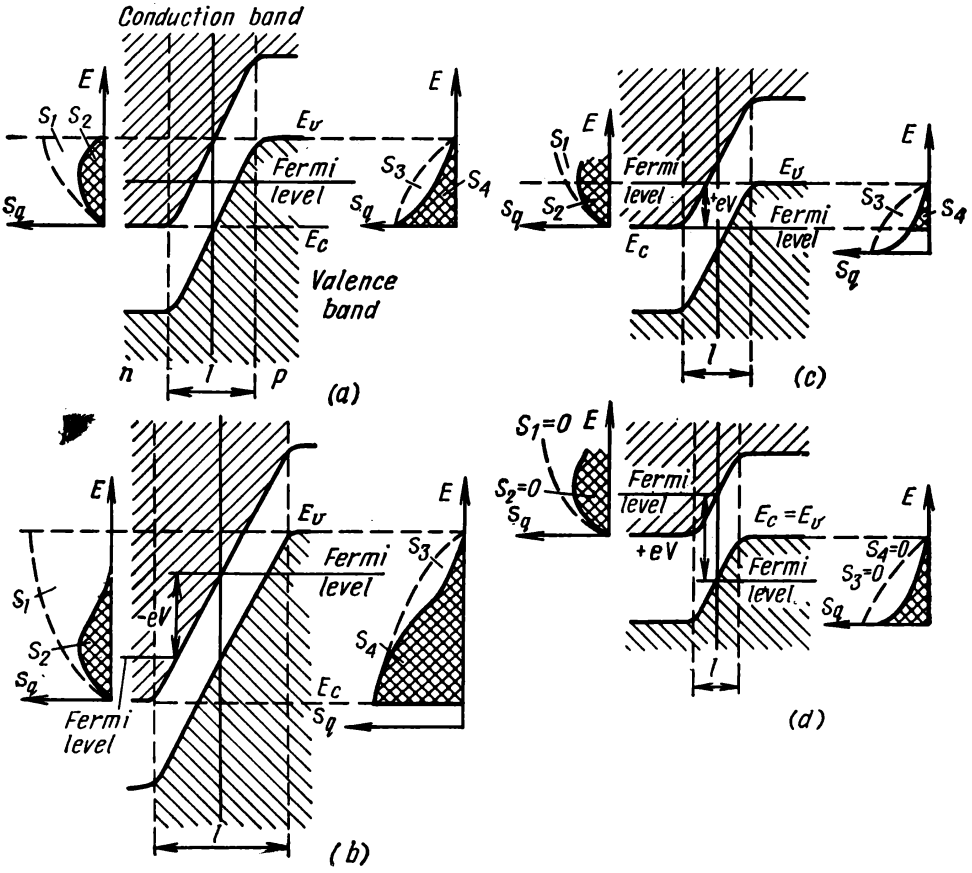


Fig. 15-13. Energy-level diagrams illustrating quantum mechanical tunnelling across a *P-N* junction

a—at equilibrium; *b*—reverse-biased; *c*—with low forward voltage; *d*—with high forward voltage

lines are proportional to the energy state densities, irrespective of their occupancy. At $T = 0$ *K*, all levels in the valence band are occupied, whereas the conduction band is empty. As a consequence, the area bounded by the dashed curve in the valence band is proportional to the occupied state density, and the area bounded by the curve in the conduction band is proportional to the unoccupied state density.

At $T > 0$ K, the electrons of the valence band move to the conduction band. The shaded areas represent the occupied state densities, whereas the remaining areas (between the dashed and solid lines) correspond to the unoccupied state densities. Of all electrons, only those whose energies lie in the interval $\Delta E = E_v - E_c$ corresponding to the band overlap, can participate in quantum mechanical tunnelling. The number of likely tunnel transitions from left to right (in the diagram) is proportional to the product of the occupied state density in the conduction band of the N-type semiconductor by the unoccupied state density in the valence band of the P-type semiconductor. This product can be obtained by multiplying together the respective areas:

$$\overrightarrow{W} = cS_2S_3 \tag{15-9}$$

where $\overrightarrow{W}$ is the number of transitions from left to right and c is the proportionality coefficient.

For the likely number of transitions in the reverse direction, we may write

$$\overleftarrow{W} = cS_4S_1 \tag{15-10}$$

At equilibrium ($V = 0$), the products of the above areas and, as a consequence, the currents in opposite directions are equal, so the resulting tunnel current is zero,

$$j_t = 0 \tag{15-11}$$

Dependence of Tunnel Current on Applied Voltage. When the junction is reverse-biased, the energy level diagram takes the form shown in Fig. 15-13*b*. The band overlap interval, ΔE, increments by $|eV|$ and encompasses the deeper occupied levels in the valence band of the P-type semiconductor, and the shallower unoccupied levels in the conductor band of the N-type semiconductor. The areas S_1 and S_4 expand more than the areas S_2 and S_3. Because of this, the number of likely tunnel transitions of electrons from the P-type into the N-type semiconductor exceeds the number of such transitions in the opposite direction:

$$S_1S_4 \gg S_2S_3 \tag{15-12}$$

$$\overleftarrow{W} \gg \overrightarrow{W} \tag{15-13}$$

Because in a P-type semiconductor, electrons are minority carriers, their predominant quantum mechanical tunnelling from the P-type to the N-type region is responsible for the rise in the reverse tunnel current, $-I_t$ (Fig. 15-14).

When the junction is forward-biased, the interval ΔE decrements by $|eV|$ (Fig. 15-13*c* and *d*). As V increases, the areas S_1, S_2, S_3 and S_4 decrease, but their relation as a function of V is more complex than for reverse biasing.

The predominant quantum mechanical tunnelling of electrons in one direction or the other is proportional to the difference between the products of the respective areas:

$$j_t = c_1 (S_2 S_3 - S_1 S_4) \qquad (15\text{-}14)$$

As V rises, S_1, S_2, S_3 and S_4 decrease by different amounts. At low values of V, S_1 and S_4 decrease at a faster rate and the difference of the products is positive, that is, $S_2 S_3 > S_1 S_4$. The quantum mechanical tunnelling of the majority carriers (electrons) from the N-type semiconductor is predominant, and the forward tunnel current increases (Fig. 15-14). At some value of V, when the Fermi level in the N-type semiconductor takes up about the same position as the top of the valence band (see Fig. 15-13*c*), the difference $S_2 S_3 - S_1 S_4$ is a maximum, and the tunnel current I_t is also a maximum. Any further rise in V is accompanied by a decrease in the difference $S_2 S_3 - S_1 S_4$ (the areas S_2 and S_3 are reduced at a faster rate), but $S_2 S_3 > S_1 S_4$ as before. Therefore, a forward tunnel current is flowing across the junction, whose value decreases with increasing V.

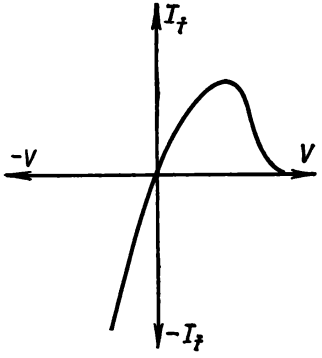


Fig. 15-14. Tunnel current as a function of applied voltage

When V becomes equal to $\Delta E/e$, the overlap interval reduces to zero (see Fig. 15-13*d*). As a consequence, $S_1 = S_2 = S_3 = S_4$, and the tunnel current drops to zero as well.

Volt-Ampere Characteristic. For a tunnel diode, the volt-ampere characteristic can be obtained by adding together the volt-ampere characteristic of a conventional P-N junction and the function $I_t = f(V)$ (see Fig. 15-15*a*). Within the reverse voltage region, the current curve is practically an exact replica of the $I_t = f(V)$ curve, and a considerably smaller reverse current, I_r, is added to the reverse tunnel current. The initial portion of the forward slope is likewise determined by the relation $I_t = f(V)$; the diffusion current is very small. At high voltages, I_t tends to zero, and the volt-ampere characteristic takes on the usual form—the current is decided by the diffusion of the majority carriers.

The plot of $I = f(V)$ in Fig. 15-15*b* shows the characteristic points serving as parameters for tunnel diodes. As the voltage varies from V_p to V_v, the device shows a negative resistance.

The Backward Diode. With a lower level of doping, the volt-ampere characteristic of a tunnel diode has the form shown in Fig. 15-16. The $I = f(V)$ curve shows no peak; because of quantum mechanical tunnelling, however, the forward branch has a more

gradual initial portion. If we turn this curve through 180°, the resulting curve will resemble the volt-ampere characteristic of a diode

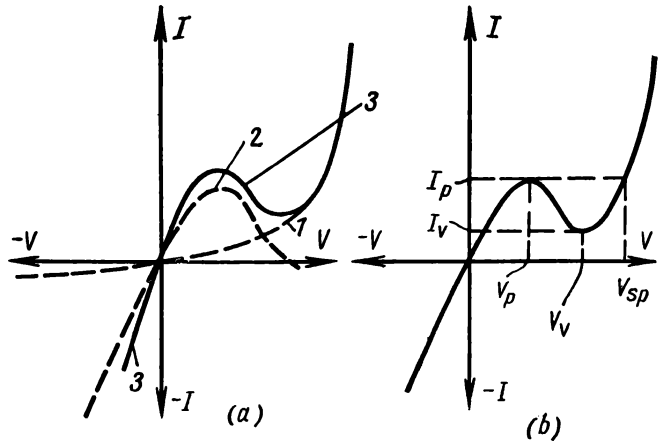


Fig. 15-15. Volt-ampere characteristics of a tunnel diode
1—diffusion current curve; *2*—tunnel current; *3*—total characteristic

with a very small resistance when forward-biased (the dashed curve). Such devices are called *backward diodes.*

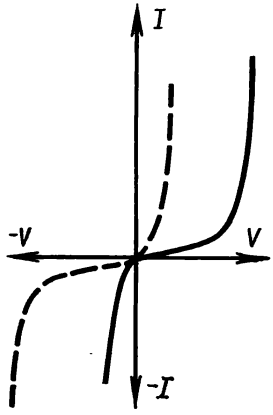


Fig. 15-16. Volt-ampere characteristic of a backward diode

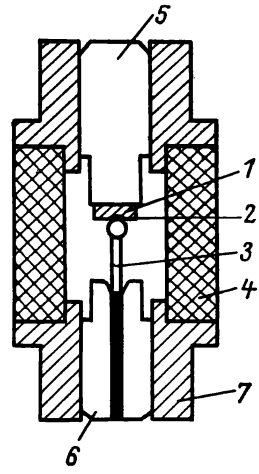


Fig. 15-17. Arrangement of a cartridge tunnel diode
1—semiconductor crystal; *2*—*P-N* junction *3*—connection; *4*—case; *5*—top lead; *6*—bottom lead; *7*—bushing

Parameters and Ratings of Tunnel Diodes. These are the voltages and currents associated with the specific points on the volt-ampere characteristic (see Fig. 15-15*b*). The peak current, I_p, corresponds to

the peak of the volt-ampere characteristic within the region of the tunnel effect. The peak voltage, V_p, is that which corresponds to the peak current. The valley current, I_v, is the current measured at the positive voltage for which the current has a minimum value. The valley voltage, V_v, is the voltage corresponding to I_v. The voltage

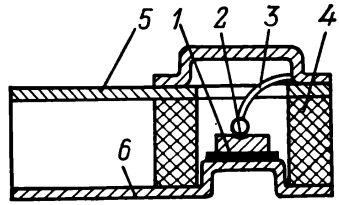


Fig. 15-18. Arrangement of a pellet tunnel diode

1—semiconductor crystal; *2*—*P-N* junction; *3*—connection; *4*—case; *5*—top lead; *6*—bottom lead

corresponding to the value of I_p on the diffusion portion of the characteristic is called the spread voltage, V_{sp}.

The falling portion of the $I = f(V)$ curve represents the negative resistance, $r_d = -dV/dI$, whose value, with a slight error, can be found by the following equation

$$r_d = (V_v - V_p)/(I_p - I_v) \qquad (15\text{-}15)$$

Sometimes, data sheets for tunnel and backward diodes specify the ratio I_p/I_v.

At present, tunnel diodes are mainly fabricated from germanium and gallium arsenide. Examples of tunnel diodes are shown in Figs. 15-17 and 15-18.

Chapter Sixteen

APPLICATIONS OF SEMICONDUCTOR DEVICES

16-1. NOISE IN SEMICONDUCTOR DEVICES

Internal noise in semiconductor devices has about the same effect on the operation of various circuits as internal noise in electron tubes (see Chap. 5). The effect of noise is especially troublesome in circuits intended to operate on weak signals, such as mixers in radio receivers, the input circuits of amplifiers, etc.

Basic Sources of Noise. In semiconductor devices, the basic sources of noise are the same as in electron tubes — fluctuations manifested as the shot effect, partition noise, thermal noise in the bulk resistance, etc. The physical aspects of these forms of noise have been examined in Sec. 5-1.

What is specific of semiconductor devices is *excess* (or low-frequency) noise caused by various surface processes.

Shot noise. As is shown in Chap. 5, the rms value of the noise current due to the shot effect is defined by Eq. (5-3) to be

$$\overline{i_n^2} = 2eI\Delta f$$

For crystal diodes, I is the current flowing through the diode. In transistors, as will be recalled, the emitter current is divided into two components, namely the current in the base circuit and the current in the collector circuit. Therefore, the rms noise current can be found for each circuit separately. Thus, for a transistor connected in a common-base circuit and operating in the active region, we may write for the emitter-base circuit:

$$\overline{i_{n,eb}^2} = 2e(1 - h_{21b})\, I_e \Delta f \tag{16-1}$$

for the collector-base circuit:

$$\overline{i_{n,bc}^2} = 2eI_{cb0}\Delta f \tag{16-2}$$

for the emitter-collector circuit:

$$\overline{i_{n,ec}^2} = 2eh_{21b}I_e\Delta f \tag{16-3}$$

Thermal noise. The rms thermal noise voltage developed by the fluctuation component across resistance R is given by Eq. (5-11) to be

$$\overline{v_n^2} = 4kTR\Delta f$$

and the rms noise current is given by Eq. (5-12) to be

$$\overline{i_n^2} = 4kT\,(1/R)\,\Delta f$$

In semiconductor devices, the sources of thermal noise may be the bulk (spreading) resistances of the emitter, r_e, base, r_b, and collector, r_c, and also the junction resistances, r_j.

For example, the rms noise current in a diode, with allowance for thermal noise in the P-N junction, may be written thus

$$\overline{i_{n,d}^2} = 2eI\Delta f + 4kT\,(1/r_j)\,\Delta f \qquad (16\text{-}4)$$

The value of noise found by the above equation agrees well with experiment for most crystal diodes.

For transistors it is usual that $r_b > r_e$ and $r_b > r_c$. So, account is mostly taken of the noise in the base bulk resistance:

$$\overline{i_{n,t,b}^2} = 4kT\,(1/r_b)\,\Delta f \qquad (16\text{-}5)$$

Excess noise. This form of noise (also known as *current noise*, *bulk noise*, or *1/f noise*) has its basic sources in fluctuations on the surface of the semiconductor crystal, fluctuations in the surface leakage current, recombinations in surface traps, and the like.

In contrast to shot and thermal noise with a fairly broad spectrum, that of excess noise is concentrated in the low-frequency region. The rms excess noise current varies with frequency, so the spectrum of this noise is limited to about 10 kHz for most devices. The frequency dependence of excess noise current can be approximated with sufficient accuracy by the following relation

$$\overline{i_{n,ex}^2} = A\,(\Delta f/f^n) \qquad (16\text{-}6)$$

where A is a coefficient usually found by experiment and equal to 10^{-16} to 10^{-18} A^2, and n is the exponent ranging between 0.9 and 1.5 for different devices.

The maximum contribution to excess noise comes from the fluctuations at P-N junctions. For the collector junction, the magnitude of excess noise is decided to a great extent by the junction width or the collector-to-base voltage, V_{cb} [24]:

$$i_{n,ex,c}^2 \approx A_c \sqrt{|V_{cb}|}\,(\Delta f/f) \qquad (16\text{-}7)$$

The magnitude of excess noise in the emitter junction is decided by the emitter current, I_e[24]:

$$\overline{i_{n,ex,e}^2} \approx A_e I_e\,(\Delta f/f) \qquad (16\text{-}8)$$

The coefficients A_c and A_e in the above expressions are in turn dependent on the collector-to-base voltage V_{cb} and the emitter current I_e, respectively.

Evaluation of Internal Noise in Semiconducor Devices. Internal noise in semiconductor devices is evaluated in terms of the same quantities as for electron tubes (see Chap. 5).

Using the concept of an *equivalent noise resistance*, it is possible to replace a "noisy" semiconductor device by some resistance, R_n, the rms noise voltage across which is equal to the magnitude of this quantity at the output of the device in question.

As an example, for a bipolar transistor the noise resistance accounting for shot and collector-junction noise may be written thus

$$R_{n,bt} \approx (e/2kT)\,(h_{21b}I_e/|Y_{21}|^2) \tag{16-9}$$

Substituting the numerical values of the constants and $T \approx 300$ K in the above equation gives

$$R_{n,bt} \approx \frac{20}{|Y_{21}|^2} h_{21b} I_e \tag{16-10}$$

For field effect transistors, the noise resistance is mainly determined by thermal noise in the channel [24]:

$$R_n \approx 0.7/g_{21} \tag{16-11}$$

where g_{21} is the conductance term of the forward transfer admittance Y_{21}, that is, the slope, S, of the curve.

The magnitude of noise in semiconductor devices may also be stated in terms of the *relative noise temperature*, t, defined by Eq. (5-18). Using this quantity, we can, for example, describe noise in the input circuit of a bipolar transistor. With allowance for shot noise, emitter-junction noise and thermal noise in the base bulk resistance, the relative noise temperature for the input conductance, g_{11}, can be written thus

$$t_{11} \approx (1/g_{11})\left[\frac{e}{2kT} I_e (1 - h_{21b}) + r_b' 4\pi f^2 C_{11}^2\right] \tag{16-12}$$

where g_{11} is the conductance term of the input admittance Y_{11} in small-signal operation, r_b' is the resistance of a part of the base between its terminal and the base-emitter junction, and C_{11} is the input capacitance of the transistor.

As with electron tubes, the key parameter used to evaluate noise in semiconductor devices is the *noise factor*, K_n. It is defined in Chapter 5, Eqs. (5-19) and (5-20).

The general relation connecting the noise factor of a bipolar transistor to its physical parameters has the form

$$K_n = 1 + (r_b/R_s) + (e/2kT)\frac{(R_s + r_b)^2}{R_s}\left[I_{cbo} + (1 - h_{21b})\,I_e + (A/2ef)\right] \tag{16-13}$$

where R_s is the internal resistance of the signal source thermal noise in which is used in accordance with Eq. (5-19) for comparison with transistor noise. In contrast to electron tubes where the input current is usually zero ($I_{g1} = 0$ at $V_{g1} < 0$), the input circuit of a transistor carries a current which is the base current, I_b, in a common-emitter circuit, or the emitter current, I_e, in a common-base current. Some of this current also divides into the signal-source circuit. Therefore, the noise factor depends on R_s.

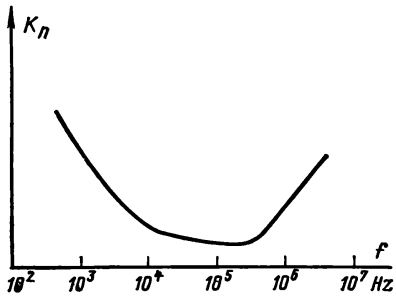


Fig. 16-1. Noise factor as a function of frequency

It is seen from Eq. (16-13) that the noise factor is frequency-dependent, which is shown in the plot of Fig. 16-1. As already noted, noise at low frequencies is essentially excess noise (the last term in the brackets). At frequencies ranging from tens to hundreds of kilohertz, the primary contribution comes from thermal fluctuations and the shot effect. As frequency rises, the current gain goes down, the base current rises, and so does shot noise owing to recombination

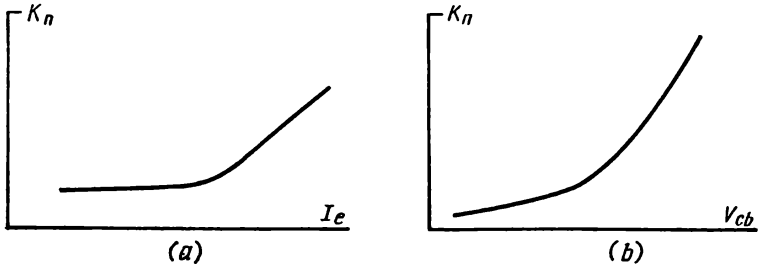


Fig. 16-2. Noise factor as a function of (*a*) emitter current and (*b*) collector-to-base voltage, V_{cb}

fluctuations. The noise factor, K_n, also depends on the emitter current, I_e, and the magnitude of the collector-to-base voltage, V_{cb} (see Fig. 16-2). At low values of I_e, thermal and excess noise is predominant, as this form of noise depends on the emitter current but little. With increasing I_e, an increasing contribution comes from shot noise. The rise in the noise factor with increasing V_{cb} is explained by the growth in excess noise.

For different transistors, the noise factor ranges from 5 to 20 dB.

16-2. LINEAR APPROXIMATION OF DIODE CHARACTERISTICS AND DIODE EQUIVALENT CIRCUITS

As with circuits using electron tubes, those using semiconductor devices are designed with the aid of equivalent circuits and linear approximation. For engineering purposes, it is most convenient to

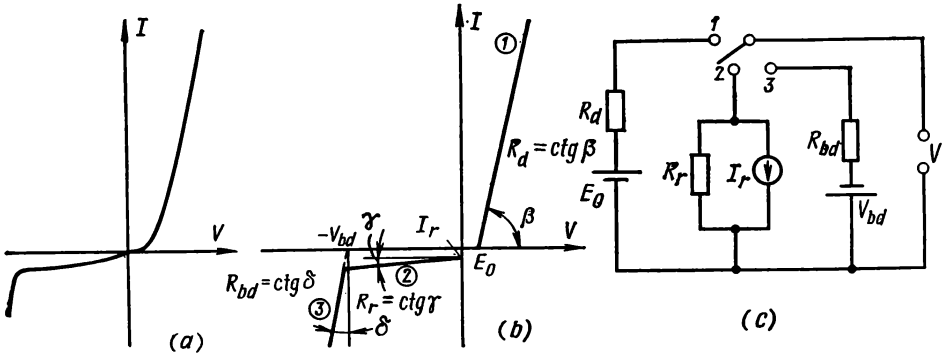


Fig. 16-3. Approximation of the characteristic of a crystal diode and its equivalent circuit

(*a*) volt-ampere characteristic of the diode; (*b*) piecewise-linear approximation of the characteristic; (*c*) equivalent circuit

use piecewise linear approximation examined with reference to electron tubes in Sec. 5-3.

The volt-ampere characteristic of a real crystal diode may be approximated by a chain of straight-line segments. Depending on the

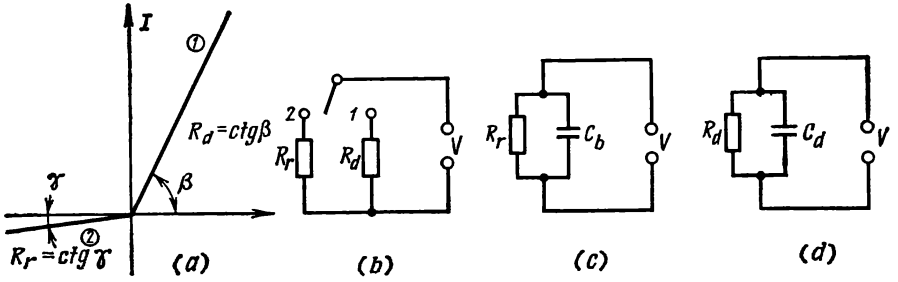


Fig. 16-4. (*a*) Piecewise-linear approximation of the diode and (*b*) through (*d*), its equivalent circuits

application of a given diode, it is usual to select and approximate the respective portion of its volt-ampere characteristic in much the same way as this is done for a vacuum diode.

The manner in which three portions of a diode characteristic can be approximated and the respective equivalent circuit are shown in

Fig. 16-3*. Another example of piecewise linear approximation and the respective equivalent circuit is shown in Fig. 16-4*a* and *b*.

As regards r.f. crystal diodes, a very important factor may be the junction capacitance, C_j. Because of this, it will be legitimate to treat the impedance of the diode, both forward and reverse biased, as a complex one. In an equivalent circuit, this is done by placing the junction capacitance, C_j, parallel with the diode resistance, R_d, and the reverse resistance, R_r (see Fig. 16-4*c* and *d*). In reverse biasing, C_j is determined by the barrier capacitance, C_{bar}; in forward biasing, it is practically equal to the diffusion capacitance, C_D.

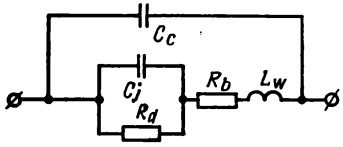


Fig. 16-5. Equivalent circuit of an r.f. diode with allowance for cartridge reactances

At high frequencies, it is often important to take into account not only the capacitance of the *P-N* junction, but also the distributed reactances associated with the structural elements of the device. For example, it is customary to extend the equivalent circuit for an r.f. diode to include the case capacitance, C_c, and also the inductance of the cat's whisker and leads, L_W (Fig. 16-5).

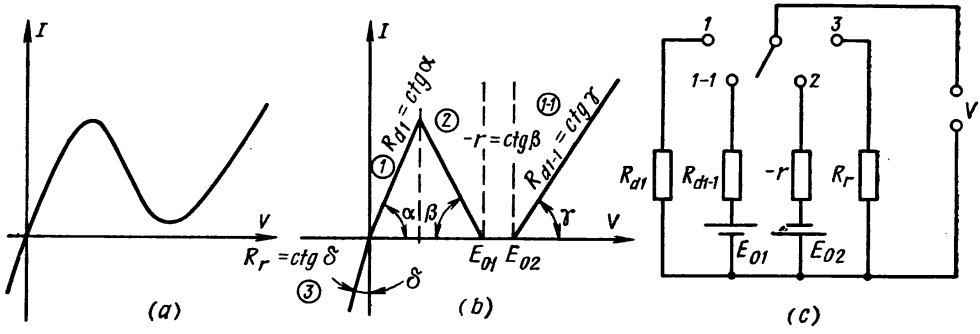


Fig. 16-6. (*a*) Volt-ampere characteristic of a tunnel diode, (*b*) its approximation, and (*c*) its equivalent circuit

An example of how the characteristic of a tunnel diode can be approximated and the respective equivalent circuit are shown in Fig. 16-6 where positions *1*, *2*, *3*, and *1-1* of the switch correspond to particular portions of the characteristic.

---

* Here and elsewhere, the definitions of incremental (dynamic) resistances in terms of trigonometric functions involve the dimensioned coefficients *m* and *n* in accordance with the dimensions of currents and voltages along the axes of the characteristics.

16-3. EQUIVALENT CIRCUITS FOR BIPOLAR TRANSISTORS

A great variety of equivalent circuits are used in the design of circuits built around transistors. The primary considerations in their choice are convenience in use, the necessity to include a particular parameter (or parameters), or the desire to standardize the design procedure.

For all the variety, however, the equivalent circuits for transistors may arbitrarily be divided into two broad classes, namely two-ports (or four-terminal networks) described by a suitable set of equations, and physical equivalent circuits depicting the physical processes taking place in a given transistor. Linear circuit theory offers a range of very simple equations with which one can readily express the parameters of any of such circuits in terms of appropriate quantities in other circuits.

Two-port Equivalent Circuits. As already noted (see Sec. 12-3), of the six sets of equations describing two-ports, those best adapted to transistors are the equations in y- and h-parameters. We shall write these equations for the alternating small-signal voltages and currents, assuming that within the range of variations in v and i the characteristics of the transistor as a two-port remain linear:

$$\left.\begin{aligned} i_1 &= y_{11}v_1 + y_{12}v_2 \\ i_2 &= y_{21}v_1 + y_{22}v_2 \end{aligned}\right\} \tag{16-14}$$

$$\left.\begin{aligned} v_1 &= h_{11}i_1 + h_{12}v_2 \\ v_2 &= h_{21}i_1 + h_{22}v_2 \end{aligned}\right\} \tag{16-15}$$

The equivalent circuits for the two-ports described by the above equations (Fig. 16-7) can readily be constructed from the following simple considerations. For example, the input current i_1 in Eqs. (16-14) is the sum of two terms, namely the current due to the input voltage v_1 across the resistor whose conductance is y_{11} (the first term) and the current due to the effect of the output voltage v_2 on the input circuit (the second term). Accordingly, the input circuit in Fig. 16-7 has two parallel branches, one containing a resistor of conductance y_{11} and an equivalent generator whose current is decided by the term $y_{12}v_2$ in the first line of Eq. (16-14), and the internal resistance is equal to infinity. A similar reasoning applies to the output section of the equivalent circuit.

There is no need to discuss in detail how the circuit of Fig. 16-7*b* is constructed, as the reader can readily do that as an exercise.

The meaning of the y- and h-parameters naturally remains the same for any circuit configuration of the transistor. The term "input" or "output", however, may be applied to the currents and voltages which are such in a given circuit configuration. For example, in a common-base circuit (Fig. 16-8*a*), $i_1 = i_e$, $i_2 = i_c$, $v_1 = v_{eb}$, and $v_2 = v_{cb}$.

The connection between the h-parameters and the physical parameters of a transistor has been shown in Sec. 12-7; $h_{11} = r_{e,d}$, $h_{12} = -1/\mu_{ce}$, $h_{21} = \alpha$, and $h_{22} = 1/r_{c,d}$. In view of these relations the equivalent circuit of Fig. 16-17*b* takes the form shown in Fig. 16-8*b*.

An equivalent two-port circuit (Fig. 16-7*a*) described by a set of equations in y-parameters can, as for electron tubes, be reduced to a circuit containing only one current generator. Such a circuit for

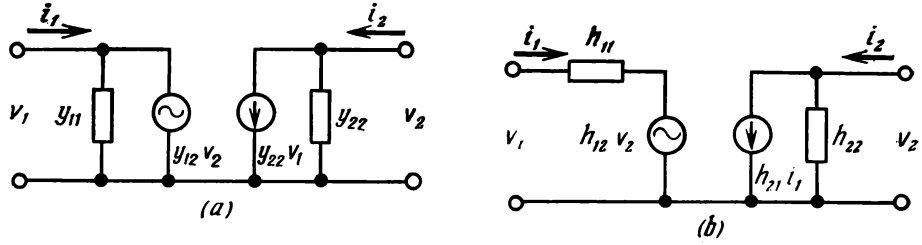


Fig. 16-7. Equivalent two-port circuit
(*a*) with y-parameter; (*b*) with h-parameters

an *N-P-N* transistor connected in a common-emitter circuit is shown in Fig. 16-9*b*. Where a transistor is used in an r.f. signal conversion circuit, the equivalent circuit must be extended to include circuit elements affecting the operation of the transistor at those frequenc-

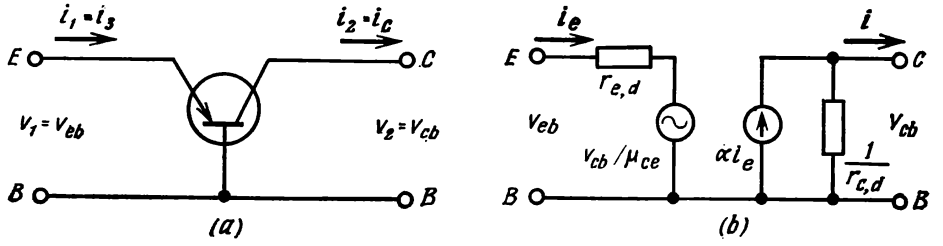


Fig. 16-8. Transistor in a common-base circuit and its equivalent circuit

ies. As an example, the equivalent circuit of Fig. 16-9*c* contains the junction capacitances and the base bulk resistance.

Physical Equivalent Circuits. An example of a physical equivalent circuit for a bipolar transistor is shown in Fig. 12-8*a*. As is noted in Sec. 12-4, this equivalent circuit applies to a transistor in any one of the four circuit configurations. Where a particular mode of operation is of interest, the circuit can readily be simplified. For example, if a transistor operates in the active region at low frequencies, the circuit can be reduced to the form shown in Fig. 16-10*a* and *b*, where the transistor is connected in a common-base and a common-emitter circuit, respectively. At high frequencies, the equivalent circuit

must be extended to include the junction capacitances, the differential junction resistances, and the equivalent generators depicting voltage feedback.

As an example, Fig. 16-10*c* shows an a.c. equivalent circuit for a transistor connected in a common-emitter circuit. The effect produc-

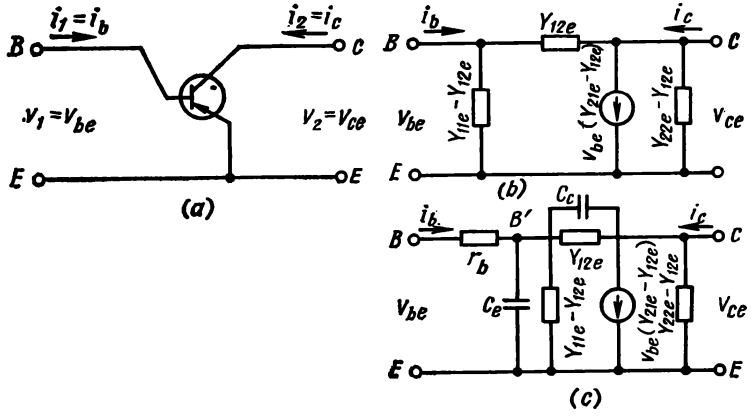


Fig. 16-9. (*a*) Transistor in a common-emitter circuit, (*b*) its equivalent circuit with one current generator, and (*c*) its equivalent circuit with allowance for junction capacitances and base resistance

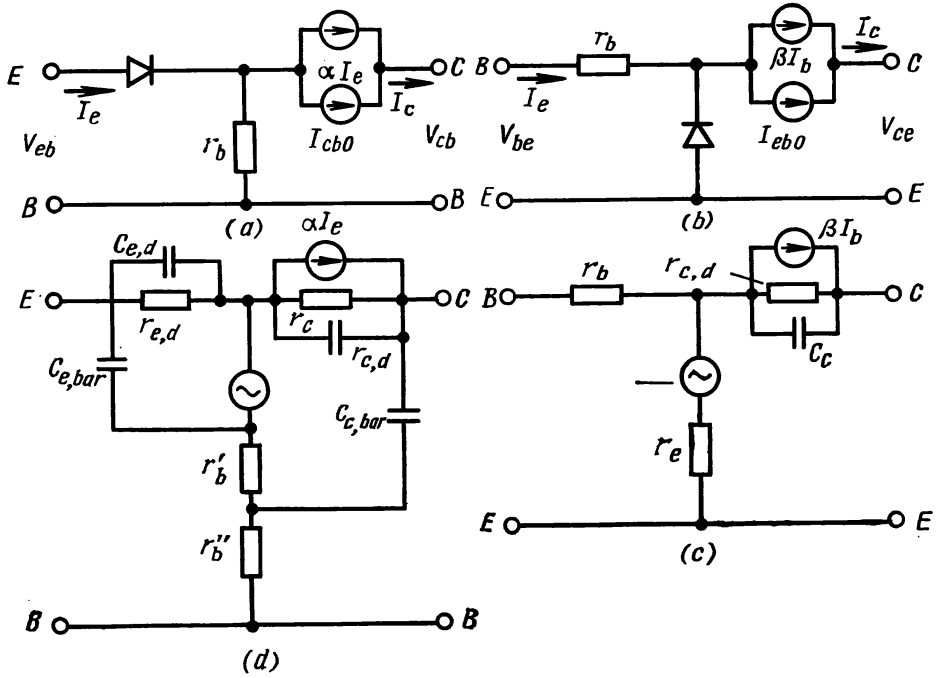


Fig. 16-10. Physical equivalent circuits of a transistor
(*a*) and (*b*) at low frequencies; (*c*) and (*d*) at high frequencies

ed by voltage feedback is represented by a generator in the emitter circuit, $| V_{ce} |/\mu_{cb}$. The collector-junction capacitance, C_c, is included in the circuit because it shunts the high-value resistance of that junction. Figure 16-10*d* shows a more elaborate equivalent circuit for a transistor connected in a common-base circuit. This circuit takes into account the diffusion and barrier capacitances and the voltage feedback (it contains a generator in the base circuit, $| V_{cb} |/\mu_{ce}$). The base resistance is represented by two equivalent resistances, r'_b and r''_b, included in the charging circuits of $C_{e,bar}$ and $C_{c,bar}$, respectively. Naturally, the parameters represented in this equivalent circuit differ in the effect they produce. For example, $r_{e,d} \ll r_{c,d}$ and the shunting effect of the emitter junction capacitances may in most cases be neglected. Therefore, the circuits of Fig. 16-10 must be regarded as examples only, and the actual equivalent circuits must be drawn up to suit the actual operating conditions of the transistor and they should only contain elements.

16-4. LINEAR APPROXIMATION OF CHARACTERISTICS FOR A BIPOLAR TRANSISTOR

The static (d.c.) characteristics of a bipolar transistor are approximated in the same manner as those of electron tubes.

Figure 16-11 shows a family of input characteristics for a transistor connected in a common-base circuit, an example of linear approximation of its characteristics, and the respective equivalent circuit. At $V_{cb} = 0$ (at the boundary between the active and saturation regions), the characteristic is approximated by a straight line intersecting the voltage axis at point E_0. Thus, it is assumed that $I_e > 0$ only when $V_{eb} > E_0$. In the equivalent circuit (the switch is in position *2*) this situation is represented by the inclusion of a battery, E_0.

In the active region, the characteristic is shifted because of voltage feedback. In the equivalent circuit (the switch is in position *1*) this situation is taken care of by inclusion of an equivalent generator, $| V_{cb} |/\mu_{ce}$.

Likely approximation of the output characteristics for a transistor in a common-base circuit is shown in Fig. 16-12. At $I_e = 0$ (at the boundary between the active and cut-off regions), the collector circuit carries a current, I_{cbo}, which is accounted for by inclusion of an equivalent current generator in position *2* of the switch. Some rise in I_c with increasing V_{cb} is represented by resistance R_2. In addition to the I_{cbo} generator and $R_1 = R_2$, the equivalent circuit for the active region contains an αI_e current generator to account for the increase in the collector current with increasing I_e. In the saturation region,

the decrease in I_c is represented by R_3 through which the current due to the positive V_{cb} flows in opposition to αI_e.

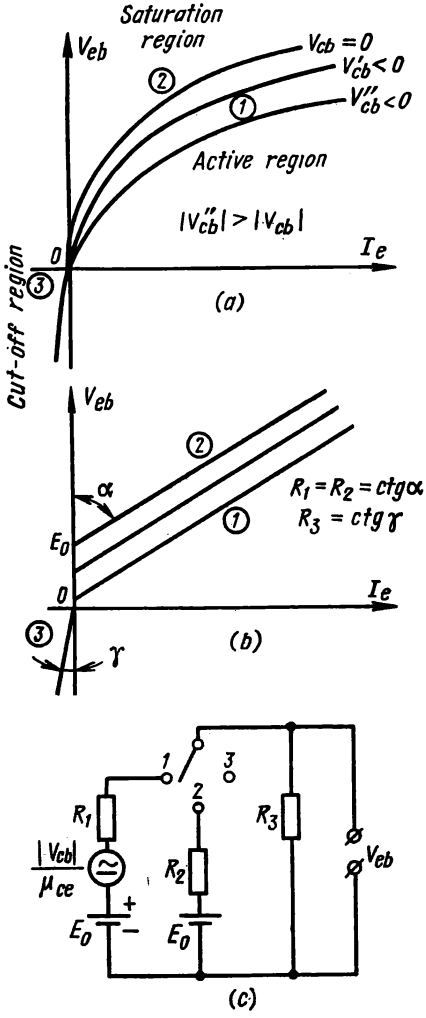


Fig. 16-11. (*a*) Real and (*b*) linearized input characteristics of a transistor in a common-base circuit, and (c) its equivalent circuit

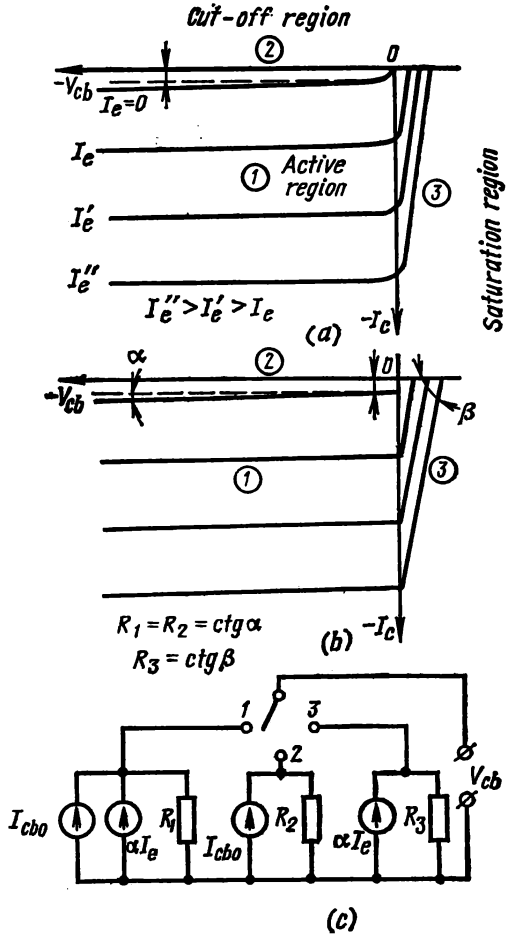


Fig. 16-12. (*a*) Real and (*b*) linearized output characteristics of a transistor in a common-base circuit, and (*c*) its equivalent circuit

16-5. EQUIVALENT CIRCUITS OF FIELD-EFFECT TRANSISTORS

Two-port equivalent circuits are also applicable to FETs. As already noted in Sec. 13-4, FETs, which are voltage-controlled devices, are best described by a set of equations in *y*-parameters defined

by Eqs. (13-14) through (13-17). A two-port equivalent circuit for a FET connected in a common-source circuit is shown in Fig. 16-13.

Figure 16-14*a* is a physical equivalent circuit for a JFET. The amplifying properties of the device are taken care of by a current generator, gV_{gs}, connected in parallel with the dynamic resistance, R_{id}. The elements R_s and R_d represent the ohmic resistances of the crystal between the source lead and the channel start and also between the drain lead and the channel end, respectively. The element

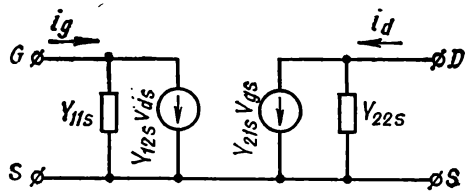


Fig. 16-13. Equivalent two-port circuit of a FET using y-parameters

R_g represents the input resistance of the device; C_g represents the barrier capacitance of the reverse-biased gate-channel junction, and R_c is the averaged resistance through which C_g charges. C_{dg}, C_{gs} and C_{ds} represent the interelectrode capacitances of the device. Because of its high value, R_g may in most cases be omitted. As a rule,

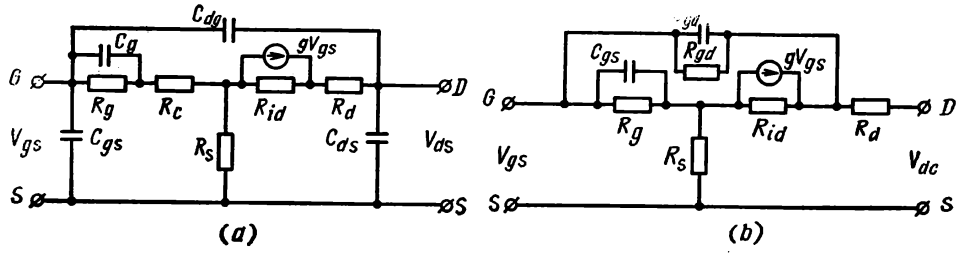


Fig. 16-14. Physical equivalent circuits of a JFET
(*a*) with one equivalent resistor, R_g; (*b*) with separate equivalent resistors, R_g and R_{gd}

R_s and R_d are likewise omitted from the equivalent circuit, because their effect is taken into account when measuring g_m and R_{id}.

Sometimes, resort is made to a modified physical equivalent circuit (Fig. 16-14*b*) where the gate resistance is represented by two elements, R_g and R_{gd}. In this circuit, C_{gs} and C_{gd} account for not only the distributed interelectrode capacitances, but also the gate barrier capacitance divided, as it were, into two parts, one at the source and the other at the drain end of the gate.

A different physical equivalent circuit is drawn up for a MOSFET (Fig. 16-15). The current generators, gV_{gs} and $g_{sub}V_{sub}$, represent the amplifying properties of the MOSFET and the ability to use I_d

in order to control both V_{gs} and V_{sub}. The element R_{id} represents the output incremental (dynamic) resistance of the device. The gate and substrate resistances are represented each by two elements; it is usual that $R_{gs} \gg R_{sub\text{-}s}$ and $R_{gd} \gg R_{sub\text{-}d}$. The gate and substrate capacitances are likewise represented each by two elements (C_{gs}, C_{gd} and $C_{sub\text{-}s}$, $C_{sub\text{-}d}$). In real MOSFETs, it is usual that $C_{gs} > C_{sub\text{-}s}$ and $C_{gd} > C_{sub\text{-}d}$.

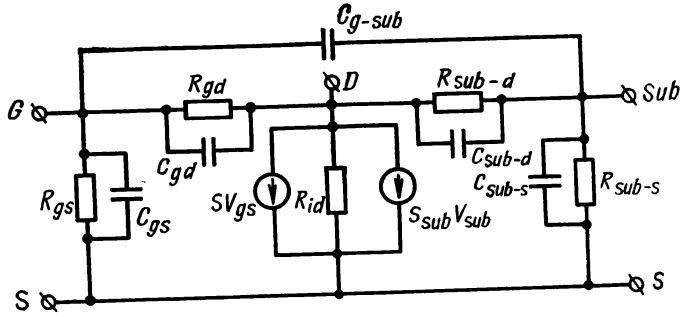


Fig. 16-15. Physical equivalent circuit of a MOSFET

16-6. RELIABILITY OF SEMICONDUCTOR DEVICES AND OPERATING VARIABLES

Reliability. The reliability of semiconductor devices can be stated in the same terms as that of electron tubes (see Sec. 5-4), namely in terms of failure rate, Eq. (5-34):

$$\lambda(t) \approx \frac{\Delta d_i}{(n-d)\,\Delta t_i}$$

and the probability of survival, Eq. (5-35):

$$P(t) \approx \frac{n - \sum_{i=1}^{t/\Delta t} \Delta d_i}{n}$$

For semiconductor devices with a sufficiently high reliability, use is often made of an approximate expression for the failure rate

$$\lambda(t) \approx d/nt \qquad (16\text{-}16)$$

The failure rate as a function of time is shown in the plot of Fig. 5-11.

Early failures in semiconductor devices mainly arise out of imperfections in manufacturing methods and hidden defects. To minimize their rate, it is customary to debug and preburn the devices after final assembly, that is, to let them operate under specified electrical conditions and elevated temperature.

Creeping (or gradual) failures which account for more than 80% of all failures occur because of a slow impairment in and deterioration of the devices, mainly caused by changes in surface physical processes under the action of temperature and humidity, thermal fatigue at the junctions, etc.

In contrast to electron tubes, it is difficult to establish the average wear time at the end of which the failure rate rapidly increases. Because of this, semiconductor devices are classed as those having a long service life. In comparison with electron tubes, semiconductor devices are simple in design and small in size. When subjected to long-term tests, many semiconductor devices have operated for several thousand hours. The failure rate of crystal diodes and transistors is about 10^{-6} to 10^{-7} per hour.

Limiting Electrical Parameters. Semiconductor devices will operate normally in a given circuit only if all the parameters of the external circuit (supply voltages, resistances, capacitances, etc.) are chosen so that the operating currents and voltages for the device are within the limits stated in its data sheet. Otherwise, the diode or transistor will fail or will show a substandard performance (giving a low gain or distorting the signal, etc.). The limits of electrical conditions for semiconductor devices are stated in terms of *maximum* and *minimum permissible values.* These are chosen with a sufficient margin over or below the values of the electrical quantities that might cause failure of the device. These maximum and minimum values are stated in reference books and data sheets.

In determining the limiting parameters, a variety of factors are taken into account, namely variations in the breakdown voltage for the collector junction that might be caused by the conditions in the emitter circuit (open-circuit, short-circuit, the value of V_{eb}), likely values of base resistance, temperature conditions, etc.

When used at their limits, semiconductor devices show a poorer performance than when they operate under normal conditions. If a device has to be used at its limits, it is important to avoid reaching the limits for two or more quantities at a time.

Thermal, Climatic and Mechanical Factors. As stated more than once, the electrical properties of semiconductors depend to a large extent on ambient temperature. In Chapters 11 and 12, it has been shown how variations in temperature can affect the basic parameters and characteristics of semiconductor devices. The working temperature range is that within which a device retains its basic parameters; it is stated between a maximum and a minimum ambient temperature, T_{max} and T_{min}. For most devices, these temperatures are +70° and —60°C, respectively. For some devices, the maximum temperature may be as high as 120°C. Some data sheets also specify the maximum temperature of the collector junction—it determines the power that a given device can dissipate.

Climatic factors include relative humidity at 40°C, maximum and minimum atmospheric pressure. In special cases, the data sheet may specify the stability of a device towards sea fog, temperature cycling, radioactivity, cosmic rays, etc.

The mechanical factors that may be stated in a data sheet are functioning and survival under conditions of vibrations at specified frequencies and *g*-load and also under conditions of single or repeated impacts.

For better heat abstraction, power diodes and transistors are fitted with external radiators (heat-sinks) attached to the case. As an alternative, the maker may recommend the manner in which a given device must be installed within an equipment for better heat withdrawal.

As a way of enhancing resistance to moisture, extremes of temperature, etc., the cases of semiconductor devices are hermetically sealed and encapsulated in suitable materials, such as silicone resins with the addition of active desiccators.

REFERENCES

1. Vlasov V. F. Vacuum and Gas-Filled Tubes. Moscow, Sviazizdat, 1960 (in Russian).
2. Khlebnikov A. N. Electron Devices. Moscow, "Sviaz" Publishers, 1964 (in Russian).
3. Batushev A. V. Electron Devices. Moscow, "Higher School" Publishers, 1969 (in Russian).
4. Katsman Yu. A. Radio- and Audio-Frequency Vacuum Tubes, Moscow, "Higher School" Publishers, 1968 (in Russian).
5. Kleiner E. Yu. Fundamentals of Vacuum Tube Theory. Moscow, "Higher School" Publishers, 1974 (in Russian).
6. Shuleikin M. V. Vacuum Tubes. Moscow, USSR Academy of Sciences' Press, 1963 (in Russian).
7. Karl R. Spangenberg. Vacuum Tubes. McGraw-Hill, 1948.
8. Gaponov V. I. Electronics. Volumes 1 and 2. Moscow, Fizmatgiz, 1960 (in Russian).
9. Tsarev B. M. Vacuum Tube Design. Moscow, "Energiya" Publisher, 1967 (in Russian).
10. Van der Ziel A. Fluctuation Noise in Radio and Physics. Russian translation. Moscow, Gosenergoizdat, 1958.
11. High-Power Microwave Tubes. Moscow, "Mir" Publishers, 1974 (in Russian).
12. Zhigarev A. A. Electron Optics and Electron-Beam Devices. Moscow, "Higher School" Publishers, 1972 (in Russian).
13. Knoll M., Kaisan B. Charge-Storage Cathode-Ray Tubes. Russian translation. Moscow, Gosenergoizdat, 1955.
14. Soboleva N. A. *et al.* Photoelectron Devices. Moscow, "Nauka" Publishers, 1965 (in Russian).
15. Erkin A. M. Cold-Cathode Tubes. Moscow, "Energiya" Publishers, 1972 (in Russian).
16. Kittel Ch. Introduction to Solid State Physics. New York, Wiley, 1953.
17. Robert L. Sproull. Modern Physics. New York, Wiley, 1956.
18. Oreshkin P. T. Semiconductor and Dielectric Physics. Moscow, "Higher School" Publishers, 1977 (in Russian).
19. Fedotov Ya. A. Fundamentals of Semiconductor Device Physics. Moscow, "Sovetskoe Radio" Publishers, 1963 (in Russian).
20. Shive J. N. The Properties, Physics and Design of Semiconductor Devices. Princeton (New York), Van Nostrand, 1959.
21. Kireyev P. S. Semiconductor Physics. Moscow, "Higher School" Publishers, 1969 (in Russian).
22. Shalimova K. V. Semiconductor Physics. Moscow, "Energiya" Publishers, 1976 (in Russian).
23. Zie S. M. Physics of Semiconductor Devices. Moscow, "Energiya" Publishers, 1973 (in Russian).
24. Stepanenko I. P. Basic Theory of Transistors and Transistor Circuits. Moscow, "Energiya" Publishers, 1973 (in Russian).
25. Pasynkov V. V., Chirkin L. K., Shinkov A. D. Semiconductor Devices. Moscow, "Higher School" Publishers, 1973 (in Russian).
26. Nosov Yu. R. Crystal Pulse Diodes. Moscow, "Sovetskoe Radio" Publishers, 1965 (in Russian).
27. Mazel E. Z. Power Transistors. Moscow, "Energiya" Publishers, 1969 (in (Russian).
28. Silicon Planar Transistors. Edited by Fedotov Ya. A., Moscow, "Sovetskoe Radio" Publishers, 1973 (in Russian).
29. Berman L. S. Introduction to Varicap Physics. Moscow, "Nauka" Publishers, 1968 (in Russian).

INDEX

Abnormal glow, 175
Absorptance, 356
Absorption, exciton, 357
  extrinsic, 358
  intrinsic, 356
Absorption factor, 134
Absorption spectrum, 198
Activator, 17
Active region, 281
A. F. current gain, 324
Alignment pin, 13
Alpha current gain, 288
Amount of modulation, 145
Amplification factor, no-load, 61
Anode, 13
Aperture angle, 147
Aperture disc, 144
Approximation, nonlinear, 125
  piecewise-linear, 125
Available noise power, 118
Avalanche breakdown voltage, 244

Band, conduction, 185
  forbidden, 193
  valence, 185
Band theory of solids, 204
Base, in vacuum tube, 13
  octal, 13
Base diagram, 14
Base spreading (bulk) resistance, 256
Base width modulation, 293
"Bath-tub" curve, 129
Beam, reading, 169
  writing, 169
Beam-confining plates, 79
Beam-forming screen, 99
Beta current gain, 289
Bias, 67
Biasing, forward, 231
  reverse, 231
Bioluminescence, 385
Blackbody, 132
Bogey values, 42
Bohr radius, 191
Bombardment-induced conductivity, 170
Boson, 205
Breakdown, avalanche, 243
  thermal, 235
  Zener, 244
Bulk noise, 404

Capacitance, barrier, 246
  cold, 41
  diffusion, 246
  dynamic, 41
  hot, 41
Capacitance ratio, 275
Capacitive reactance, 42
Carrier collection coefficient, 366
Carriers, equilibrium, 199
  majority, 196
  minority, 196
  non-equilibrium, 200
Cathode, 13
  barium-oxide, 26
  boride, 25
  cold, 13
  directly heated, 22
  filament, 22
  filament characteristic of, 23
  heater, 22
  hot, 13
  indirectly heated, 22
  oxide-grid, 26
  secondary-emission, 13
  thermionic, 13
  thorium-oxide, 26
  tungsten, 24
Cathode-fall space, 175
Cathode screen, 97
Cathodoluminescence, 159, 385
Catwhisker, 250
Channel, 339
Charge density, 33
Charge image, 168
Chemical potential, 204
Chemiluminescence, 385
Collection efficiency, 139
Collector efficiency factor, 289
Complete (sudden) failures, 129
Composite controlling voltage, 46
  in multielectrode tubes, 81
Composite current, 27
Concentration gradient, 200
Conduction, electron (N-type), 189
  extrinsic, 219
  hole (P-type), 189
  intrinsic, 219
Conductivity, 218
Conductor, 185
Configuration, common-base, 280
  common-cathode, 45
  common-collector, 280
  common-drain, 341
  common-emitter, 280
  common-gate, 341
  common-grid, 45
  common-plate, 45
  common-source, 341
Continuity equation, 223
Conversion loss, 272
Conversion parameters, 110
Critical potential, first, 159
  second, 160
Crossover, 147
CRT, 141
  barrier-grid storage, 168
  character-display, 165
  dark-trace, 164

electrostatic, 141
magnetic, 141
oscilloscopic, 162
double-beam, 163
dual-trace, 163
postacceleration, 163
radar, 164
scan-conversion, 169
shaped-beam, 165
storage, 167
subtracting, 169
Current partition, 51
in multielectrode tubes, 83
Current partition ratio, 51
Cut-off characteristic, remote, 88
sharp, 87
Cut-off frequency, alpha, 324
beta, 324
Cut-off region, 281
Cut-off voltage, 46

Dark conductivity, 359
Dark current, 136
Dark resistance, 363
Decade counter, double-pulse, 182
Deflection coil, 143
Deflection plates, bent, 156
curved, 156
diverging, 156
Deflection sensitivity, 155
Deflection structure, 142
Deflection yoke, 165
Delay time, 323
Density current, 18
Depletion (barrier) layer, 229
Diffusion constant, 221
Diffusion current, 221
Diffusion current density, 222
Diffusion length, 224
Diffusion velocity, 224
Dinistor, 391
Diode, crystal, 249
avalanche, 390
backward, 400
capacitances of, 262
classification of, 250
detector, 271
double-base, 390
for pulse work, 268
Gunn, 390
junction type, 250
mixer, 271
parameters of, 261
point-contact type, 250
snap-off (charge-storage), 270
Schottky, 270
transients in, 263
tunnel, 396
Zener, 273
vacuum, 11, 29
equivalent circuit for, 121
ideal, characteristics of, 35
interelectrode capacitance of, 40
parameters of, 39
potential distribution in, 30
principle of operation of, 29
space-charge-limited operation of, 31
Dipole layer, 16
Discharge, arc, 175
dark, 137
glow, 174
high-frequency, 176
nonself-maintained, 174
self-maintained, 174
Townsend, 174
trigger, 181
Displacement current, 28
Distributed inductances, 103
Distributed reactances, 103
Double-grid current control, 91
Double-grid current control factor, 91
Drain, 339
Drift current, 207
Drift mobility, 217
Dynamic alpha current gain, 288
Dynamic beta current gain, 290
Dynatron, 98
Dynatron effect, 78
Dynode, 13, 98
squirrel-cage, 140
Dynode current, 98
Dynode voltage, 99

Early effect, 293
Einstein photoelectric equation, 133
Einstein relationship between mobility and diffusion constant, 222
Effective plate area, 34
Electric fluctuations, 115
Electrochemical potential, 207
Electroluminescence, 385
Electron, primary, 20
secondary, 20
Electron-beam device, 141
Electron-beam photoelectric device, 132
Electron burn, 160
Electron drift, 216
Electron flyback, 51
Electron gas, 205
degenerate, 208
non-degenerate, 208
Electron gun, 142
reading, 169
tetrode, 151
triode, 149
writing, 169
Electron-hole pair, 194
Electronic work function, 133
Electron interception, 52
Emission, electron, 17
field, 17
photoelectric, 17
secondary, 12
thermionic, 13
Emission current density, 18
Emission efficiency, 22
Emitter, atomic-film, 24
bariated-tungsten, 24
oxide-coated, 24
matrix, 24
pure metal, 24
secondary-electron, 20
thoriated-tungsten, 24
Energy barrier, 16
Energy distribution function, 17, 207
Energy gap, 185
Energy level, 16
Energy state density function, 205
Equivalent circuit, a.c., 129
for high frequencies, 123
Equivalent noise resistance, 117
Excess noise, 403
Excitation voltage, 172
Extraction of minority carriers, 235

Failure rate, 129
Family of characteristics, 16
Feedback, 76

Fermi-Dirac statistics, 17, 205
Fermi-Dirac distribution function, 206
Fermi energy, 16
Fermion, 205
FET, 339
  N-channel, 339
  parameters of, 352
  P-channel, 339
  principle of operation of, 341
Field, accelerating, 19
    retarding, 16
Flicker effect, 117
Fluctuation voltage, 115
Fluorescence, 159
Focal distance, 149
Focusing coil, 143
Focusing structure, 146
Frequency, carrier, 108
  intermediate, 108

Gas-amplification factor, 137
Gas current, 57
Gate, 339
Generation, field, 194
  impact, 194
  light, 194
    bipolar, 198
    monopolar, 198
  rate of, 199
  thermal, 194
Generation current, 253
Grid, 13
  barrier, 168
  frame, 74
  modulating, 141
  punched rod, 74
  screen, 77
Grid leakage current, 130
Ground state energy, 191
Guide pulse, 182

Heptode, 108
Hole, 194
Horizontal-deflection plates, 142
"Hot" electron, 200

IGFET, 339
Ignitron, 20, 171
Image charge, 16
Impurity, donor, 189
  acceptor, 189
Impurity centres, 202
Impurity energy band, 192
Induced charge, 27
Induced current, 26
Injection level, 233
Injection of minority carriers, 232
Injection ratio, 255
Input chamber, 139
Insulator, 185
Intensifier electrodes, 163
Intensity modulation, 143
Inverse current gain, 288
Inversion region, 241
Ionization, surface, 173
  volume, 173
Ionization voltage, 172
Island effect, 56

JFET, 339
Junction,
  alloy, 226
  diffused, 226
  heterogeneous, 226
  metal-semiconductor, 226
    metal-N-type semiconductor, 239
    metal-P-type semiconductor, 241
  N-I, 239
  N-N, 226
  ohmic, 242
  P-I, 239
  P-N, 226
    abrupt, 227
    asymmetrical, 238
    breakdown of, 243
    capacitances of, 245
    graded, 238
    symmetrical, 226
    unipolar, 238
    volt-farad characteristic of, 247
  P-P, 226
  Schottky, 241

Kinescope, 166
Kirchhoff's voltage law, 37

Lagrange-Helmholtz equation, 147
Lattice constant, 192
L-cathode, dispenser-type, 24
  impregnated type 25
Leakage current, 254
Lens, aperture, 146
  condenser, 147
  electron, 142
  focusing, 143
  immersion objective, 146
  immersion projective, 146
  unipotential, 146
Level, excitation, 189
  ionization, 189
  shallow, 191
Life of cathode, 22
Lifetime of free charge carriers, 199
Light emitting capacitor, 386
Light-emitting diode, 387
Limiting frequency, 363
Limiting potential, 160
Load, 15
Load impedance, 97
Load line, 37
Local diffusion, 278
Lower gate, 351
Low-voltage arc, 175
Luminescence, 385
Luminous characteristic, 135
Luminous efficiency, 132
Luminous efficiency of phosphor, 160
Luminous flux, 132
Luminous intensity, 133

Maintaining voltage, 179
Maximum deflection angle, 156
Maximum frequency of oscillation, 325
Maxwell-Boltzmann statistics, 204
Mean free path, 216
Mean free time, 216
Mean thermal velocity, 216
Megatron, 107
Mercury rectifier, 171
Metastable state, 172
Minimum discernible signal, 115
Mixer, 108
Mode, dynamic, 15
  quasistatic, 15
  static, 15

Modulation characteristic, 144
Modulation of base resistance, 256
MOSFET, depletion-mode, 339
  enhancement-mode, 339
Multiplication factor, 244

Negative resistance device, 389
Noise factor, 120
Noise figure, 120
Noise temperature, 272
Nonlinearity factor, 275
Nonoperating region, 127
N-type device, 389
N-type response curve, 389
Numerical readout indicator, 387
Nuvistor, 114
Nyquist equation, 118

Operating conditions, climatic, 15
  electrical, 15
  mechanical, 15
Operating region, 93
Operating temperature of cathode, 21
Output circuit reaction, 314

Parameter devices, 275
Partial failures, 130
Particle statistics, 204
Partition noise, 117
Path of operation, 68
Pauli's exclusion principle, 205
Peak pulse power, 272
Penetration factor, 49
Pentode, 76
Permittivity of free space, 33
Persistence, 159
Perveance, 34
Phonon, 194
Phosphor, 158
Phosphorescence, 160
Photocathode, 12, 133
Photocell, 131
Photocell, parameters of, 372
  semiconductor, 131
Photoconductive effect, 131, 355
Photoconductivity, 359
Photocurrent, 362
Photocurrent gain, 384
Photodiode, 131, 355
  avalanche, 379
  parameters of, 377
  P-I-N, 378
  Schottky-barrier, 379
  with nonuniformly doped base, 377
Photoelectric emission, 133
Photo emf, 366
Photoemission, normal, 134
  selective, 134
Photoemissive effect, 131
Photoluminescence, 385
Photomultiplication, 380
Photoresistive cell, 355
Photoresistor, 131, 355
  characteristics and parameters of, 362
Photothyristor, 355
Phototransistor, 131, 355
  parameters of, 384
  principle of operation of, 382
Phototube, 12, 131
  gas-filled, 131
  principle of operation of, 131
  vacuum, 131
Photovoltaic cell, 355
Picture definition, 166
Pinch-off voltage, 342
Pin connection diagram, 14
Plasma, 173
Plate dissipation, 66
Plate reaction, 70
P-N boundary, 227
P-N transition region, 186
Poisson's equation, 33, 230
Potential barrier, 17
Potential control, 181
Potential gradient, 143
Power gain, 325
Power-to-mass ratio, 374
Preaccelerating electrode, 145
Probability of survival, 129
Pulse resistance, 270
Punch-through, 309

Quality factor, 276
Quantum particle statistics, 205
Quasimomentum, 201
Quasi-neutral region, 233
Quiescent point, 63

Radiant energy, 132
Radiant flux, 132
Radioluminescence, 385
Recombination, 172
  emissive, 385
  interband (band-to-band), 194
  nonradiative, 194
  radiative, 194
  rate of, 199
  through local impurity levels, 194
Recombination centre, 158
Recombination current, 255
Recombination factor, 199
Recombination trap, 194
Rectifier diode, 267
Reflectance, 355
Reflection coefficient, 371
Refractive index, 147
Relative noise temperature, 120
Relaxation of photoconductivity, 360
Reliability, 129
Retrace, 163
Reverse current, 244
Reverse saturation current, 235
Ripple, 22

Saturation, 31
Saturation factor, 311
Saturation impurity-depletion region, 214
Saturation photocurrent, 133
Saturation region, 281
Saturation resistance, 311
Saturation temperature, 214
Scan, 163
Scattering centre, 216
Schottky barrier, 270
Schottky effect, 19
Schrot and Schottky noise, 115
Screen, 143
  aluminized, 161
  composite, 161
  double-layer, 161
Secondary-electron yield, 21
Secondary-emission ratio, 21
Selection plates, 165
Self-discharge, 137
Semiconductor, 185
  energy diagram of, 189

extrinsic, 196
intrinsic, 189
N-type, 196
P-type, 196
Semiconductor device, basic classification of, 188
definition of, 185
reliability of, 415
Semimetals, 193
Sensitivity, current, 273
dynamic, 363
illumination, 363
integrated, 135, 363
integrated specific luminous, 363
luminous, 363
monochromatic, 363
radiant, 363
relative, 132
static, 363
voltage, 273
Shake-down period, 129
Shot noise, 115
Solar battery, 355
Solar cell, 131
Source, 339
Space charge, 30
Space-charge noise-reduction factor, 116
Spectral characteristic, 134
Spectral photoconductivity characteristic, 360
Spread voltage, 402
Stage efficiency, 139
Stage gain, 97
Stefan-Boltzmann law, 43
Stimulated emission, 385
S-type device, 389
S-type response curve, 389
Substrate, 188
Susceptance, 322
Sweep voltage, 163
Switching charge, 269

Target, 20, 164
Temperature coefficient of capacitance, 277
of photocurrent, 364
of Q-factor, 277
of voltage, 261
Tetrode, 76
characteristics of, 77
beam-power, 79
Thermal-agitation noise voltage, 118
Thermal breakdown voltage, 245
Thermal current, 235
Three-halves power law, 32
for vacuum triode, 49
Threshold flux, 363
Threshold frequency, 134
Threshold of photoemission, 133
Threshold voltage, 349
Thyratron, 171
breakdown characteristic of, 181
control characteristic of, 177
current controlled, 181
firing characteristic of, 181
glow, 179
high-voltage, 176
low-voltage, 176
pulse, 176
screen-grid, 178
tetrode, 180
triode, 180
Thyristor, parameters of, 395
Total current density, 222
Total reverse current, 254
Transconductance, 60
dynamic, 71
reverse, 63
Transfer current, 101
Transistor, alloy, 333
alloy-diffused, 337
bipolar, 188, 278
definition of, 278
diffusion, 335
drift, 335
high-frequency high-power, 337
high-frequency low-power, 333
high-frequency parameters of, 323
high-power, 337
h-parameters of, 297
low-frequency low-power, 333
N-P-N, 280
planar, 278
P-N-P, 280
pulse operation of, 320
pulse parameters of, 329
r.f. operation of, 320
small-signal parameters of, 290
static parameters of, 310
unijunction, 389
unipolar, 339
y-parameters of, 297
Transit angle, 101
Transition, direct, 201
downward, 384
indirect, 201
radiative, 385
Transit time, 100
Transit-time effects, 100
Triode, vacuum, 11
equivalent circuit for, 122
interelectrode capacitances of, 66
lighthouse, 106
metal-ceramic, 106
static characteristics of, 53
static parameters of, 59
penetration of, 48
potential distribution in, 45
Trinistor, 391
Tube, amplifying, 11
arc-discharge, 12, 176
decade counting, 182
disc-seal, 107
double-beam, 11
dual, 108
efficiency of, 112
electrometer, 114
electron, base-type, 13
figure of merit of, 75
frequency-converter, 11, 108
gas-filled, 11, 171
gas rectifier, 12
generating, 11
glow-discharge, 12, 179
high-frequency-discharge, 179
high-power, 111
mercury-pool rectifier, 20
microwave, 105
multibeam, 11
multielectrode, 11
interelectrode capacitances of, 94
parameters of, 89
multisection, 107
photomultiplier, 131
current gain of, 139
radio-frequency, 100
rectifying, 11
screened-cathode, 97
secondary-emission, 98
single-beam, 11
soft-lead, 14
spark-gap, 171

stabilizer, 117
  stiff-lead, 14
  TV camera, 12
  vacuum, 11
  VR, 12
Tube characteristic, static, 15
  dynamic, 15
Tube equation, 62
Tube noise, 115
Tunnel current, 244
Tunnel effect, 20, 134
Turn-off time, 269
Turn-on time, 269
Two-port, 59

Ultor, 167
Unit bandwidth threshold flux, 364
Useful-life period, 129

Valence electron, 189
Valley current, 402
Valley voltage, 402
Varactor, 275
Varicap, 275
Vertical-deflection plates, 143
Virtual cathode, 80
Voltage amplification factor, dynamic, 72
Voltage feedback, 294
Voltage feedback factor, 294
Voltage stage gain, 315

Wave function, 192
Wave number, 201
Wear-in period, 129
Wear-out period, 129
Work function, 16

Y-parameters, 59

To the Reader

Mir Publishers welcome your comments on the content, translation, and design of the book.

We would also be pleased to receive any suggestions you care to make about our future publications.

Our address is:
USSR, 129820,
Moscow, I-110, GSP,
Pervy Rizhsky Pereulok, 2,
Mir Publishers

Printed in the Union of Soviet Socialist Republics

This English translation of the book (which has seen three editions in Russian) by one of the leading Soviet experts on electronics is a comprehensive survey of all the physical laws which contribute to the understanding of electron device behaviour.

Here, the author presents a unified fundamental treatment of electron devices including vacuum tubes and transistors, with emphasis on their similar features, rather than their differences, through the common denominator of semiconductor theory.

The book stresses the internal physics of the devices and discusses the role that physics plays in determining the external characteristics.

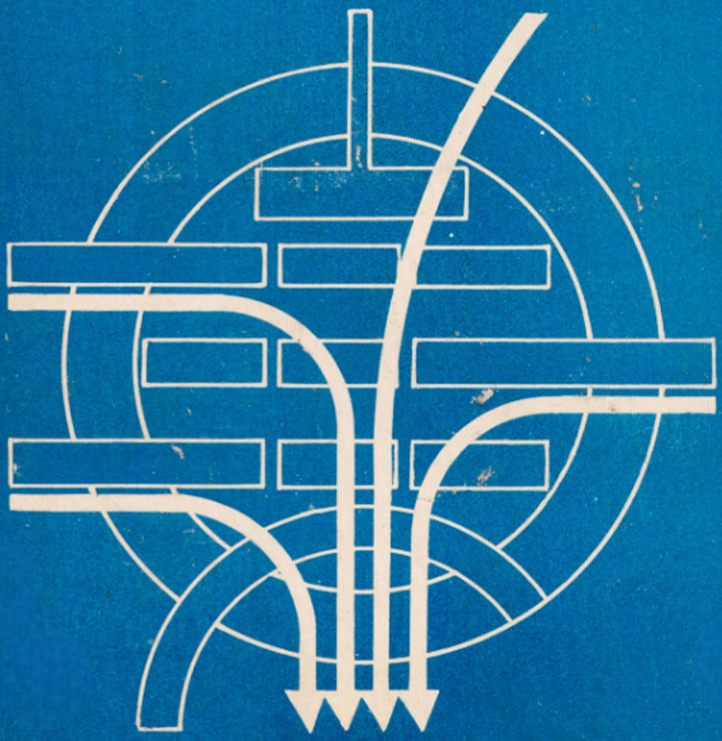
A new approach is represented in the treatment of energy levels in semiconductors and the influence of these on junction effects, emission, and photoeffects.

Recent advances include the latest forms of junction transistors, various types of FETs, and photoconductive and photovoltaic cells.

The outstanding types of devices are covered individually, and the essentials of the subject are presented in as straightforward and elementary a way as is consistent with a comprehensive treatment.

V. Dulin

ELECTRON DEVICES



Mir Publishers Moscow